Springer Series in Statistics

Michael R. Kosorok

# Introduction to Empirical Processes and Semiparametric Inference

Springer

# Springer Series in Statistics

# Springer Series in Statistics

*Alho/Spencer:* Statistical Demography and Forecasting
*Andersen/Borgan/Gill/Keiding:* Statistical Models Based on Counting Processes
*Atkinson/Riani:* Robust Diagnostic Regression Analysis
*Atkinson/Riani/Ceriloi:* Exploring Multivariate Data with the Forward Search
*Berger:* Statistical Decision Theory and Bayesian Analysis, 2nd edition
*Borg/Groenen:* Modern Multidimensional Scaling: Theory and Applications, 2nd edition
*Brockwell/Davis:* Time Series: Theory and Methods, 2nd edition
*Bucklew:* Introduction to Rare Event Simulation
*Cappé/Moulines/Rydén:* Inference in Hidden Markov Models
*Chan/Tong: Chaos:* A Statistical Perspective
*Chen/Shao/Ibrahim:* Monte Carlo Methods in Bayesian Computation
*Coles:* An Introduction to Statistical Modeling of Extreme Values
*Devroye/Lugosi:* Combinatorial Methods in Density Estimation
*Diggle/Ribeiro:* Model-based Geostatistics
*Dudoit/Van der Laan:* Multiple Testing Procedures with Applications to Genomics
*Efromovich:* Nonparametric Curve Estimation: Methods, Theory, and Applications
*Eggermont/LaRiccia:* Maximum Penalized Likelihood Estimation, Volume I: Density Estimation
*Fahrmeir/Tutz:* Multivariate Statistical Modeling Based on Generalized Linear Models, 2nd edition
*Fan/Yao:* Nonlinear Time Series: Nonparametric and Parametric Methods
*Ferraty/Vieu:* Nonparametric Functional Data Analysis: Theory and Practice
*Ferreira/Lee:* Multiscale Modeling: A Bayesian Perspective
*Fienberg/Hoaglin:* Selected Papers of Frederick Mosteller
*Frühwirth-Schnatter:* Finite Mixture and Markov Switching Models
*Ghosh/Ramamoorthi:* Bayesian Nonparametrics
*Glaz/Naus/Wallenstein:* Scan Statistics
*Good:* Permutation Tests: Parametric and Bootstrap Tests of Hypotheses, 3rd edition
*Gouriéroux:* ARCH Models and Financial Applications
*Gu:* Smoothing Spline ANOVA Models
*Gyöfi/Kohler/Krzyźak/Walk:* A Distribution-Free Theory of Nonparametric Regression
*Haberman:* Advanced Statistics, Volume I: Description of Populations
*Hall:* The Bootstrap and Edgeworth Expansion
*Härdle:* Smoothing Techniques: With Implementation in S
*Harrell:* Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis
*Hart:* Nonparametric Smoothing and Lack-of-Fit Tests
*Hastie/Tibshirani/Friedman:* The Elements of Statistical Learning: Data Mining, Inference, and Prediction
*Hedayat/Sloane/Stufken:* Orthogonal Arrays: Theory and Applications
*Heyde:* Quasi-Likelihood and its Application: A General Approach to Optimal Parameter Estimation
*Huet/Bouvier/Poursat/Jolivet:* Statistical Tools for Nonlinear Regression: A Practical Guide with S-PLUS and R Examples, 2nd edition
*Ibrahim/Chen/Sinha:* Bayesian Survival Analysis
*Jiang:* Linear and Generalized Linear Mixed Models and Their Applications
*Jolliffe:* Principal Component Analysis, 2nd edition
*Knottnerus:* Sample Survey Theory: Some Pythagorean Perspectives
*Konishi/Kitagawa:* Information Criteria and Statistical Modeling

*(continued after index)*

Michael R. Kosorok

# Introduction to Empirical Processes and Semiparametric Inference

Springer

Michael R. Kosorok
Department of Biostatistics
University of North Carolina
3101 McGavran-Greenberg Hall
Chapel Hill, NC 27599-7420
USA
kosorok@unc.edu

ISBN 978-0-387-74977-8 e-ISBN 978-0-387-74978-5

Library of Congress Control Number: 2007940955



Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

*To*
*My Parents, John and Eleanor*
*My Wife, Pamela*
*My Daughters, Jessica and Jeanette*
*and*
*My Brothers, David, Andy and Matt*

# Preface

The goal of this book is to introduce statisticians, and other researchers with a background in mathematical statistics, to empirical processes and semiparametric inference. These powerful research techniques are surprisingly useful for studying large sample properties of statistical estimates from realistically complex models as well as for developing new and improved approaches to statistical inference.

This book is more of a textbook than a research monograph, although a number of new results are presented. The level of the book is more introductory than the seminal work of van der Vaart and Wellner (1996). In fact, another purpose of this work is to help readers prepare for the mathematically advanced van der Vaart and Wellner text, as well as for the semiparametric inference work of Bickel, Klaassen, Ritov and Wellner (1997). These two books, along with Pollard (1990) and Chapters 19 and 25 of van der Vaart (1998), formulate a very complete and successful elucidation of modern empirical process methods. The present book owes much by the way of inspiration, concept, and notation to these previous works. What is perhaps new is the gradual—yet rigorous—and unified way this book introduces the reader to the field.

The book consists of three parts. The first part is an overview that concisely covers the basic concepts in both empirical processes and semiparametric inference, while avoiding many technicalities. The second part is devoted to empirical processes, while the third part is devoted to semiparametric efficiency and inference. In each of the last two parts, the chapter following the introductory chapter is devoted to the relevant mathematical concepts and technical background needed for the remainder of the

part. For example, an overview of metric spaces—which are necessary to the study of weak convergence—is included in Chapter 6. Thus the book is largely self contained. In addition, a chapter devoted to case studies is included at the end of each of the three parts of the book. These case studies explore in detail practical examples that illustrate applications of theoretical concepts.

The impetus for this work came from a course the author gave in the Department of Statistics at the University of Wisconsin-Madison, during the Spring semester of 2001. Accordingly, the book is designed to be used as a text in a one- or two-semester sequence in empirical processes and semiparametric inference. In a one-semester course, most of Chapters 1–10 and 12–18 can be covered, along with Sections 19.1 and 19.2 and parts of Chapter 22. Parts of Chapters 3 and 4 may need to be skipped or glossed over and other content judiciously omitted in order fit everything in. In a two semester course, one can spend the first semester focusing on empirical processes and the second semester focusing more on semiparametric methods. In the first semester, Chapters 1 and 2, Sections 4.1–4.3, Chapters 5–10, 12–14 and parts of Chapter 15 could be covered, while in the second semester, Chapter 3, Sections 4.4–4.5, the remainder of Chapter 15, and Chapters 16–22 could be covered in some detail. The instructor can utilize those parts of Chapter 11 and elsewhere as deemed appropriate. It is good to pick and choose what is covered within every chapter presented, so that the students are not given too much material to digest.

The books can also be used for self-study and can be pursued in a basically linear format, with the reader omitting deeper concepts the first time through. For some sections, such as with Chapter 3, it is worth skimming through to get an outline of the main ideas first without worrying about verifying the math. In general, this kind of material is learned best when homework problems are attempted. Students should generally have had at least half a year of graduate level probability as well as a year of graduate level mathematical statistics before working through this material.

Some of the research components presented in this book were partially supported by National Institutes of Health Grant CA075142. The author thanks the editor, John Kimmel, and numerous anonymous referees who provided much needed guidance throughout the process of writing. Thanks also go to numerous colleagues and students who provided helpful feedback and corrections on many levels. A partial list of such individuals includes Moulinath Banerjee, Hongyuan Cao, Guang Cheng, Sang-Hoon Cho, Kai Ding, Jason Fine, Minjung Kwak, Bee Leng Lee, Shuangge Ma, Rajat Mukherjee, Andrea Rotnitzky, Rui Song, Anand Vidyashankar, Jon Wellner, Donglin Zeng, and Songfeng Zheng.

Chapel Hill
October 2007

# Contents

# Part I

# Overview

# 1
# Introduction

Both empirical processes and semiparametric inference techniques have become increasingly important tools for solving statistical estimation and inference problems. These tools are particularly important when the statistical model for the data at hand is *semiparametric*, in that it has one or more unknown component that is a function, measure or some other infinite dimensional quantity. Semiparametric models also typically have one or more finite-dimensional Euclidean parameters of particular interest. The term *nonparametric* is often reserved for semiparametric models with no Euclidean parameters. Empirical process methods are powerful techniques for evaluating the large sample properties of estimators based on semiparametric models, including consistency, distributional convergence, and validity of the bootstrap. Semiparametric inference tools complement empirical process methods by, among other things, evaluating whether estimators make efficient use of the data.

Consider the semiparametric model

$$Y = \beta' Z + e, \tag{1.1}$$

where $\beta, Z \in \mathbb{R}^p$ are restricted to bounded sets, prime denotes transpose, $(Y, Z)$ are the observed data, $\mathrm{E}[e|Z] = 0$ and $\mathrm{E}[e^2|Z] \leq K < \infty$ almost surely, and $\mathrm{E}[ZZ']$ is positive definite. Given an independent and identically distributed (i.i.d.) sample of such data $(Y_i, Z_i), i = 1 \ldots n$, we are interested in estimating $\beta$ without having to further specify the joint distribution of $(e, Z)$. This is a very simple semiparametric model, and we definitely do

not need empirical process methods to verify that

$$\hat{\beta} = \left[\sum_{i=1}^{n} Z_i Z_i'\right]^{-1} \sum_{i=1}^{n} Z_i Y_i$$

is consistent for $\beta$ and that $\sqrt{n}(\hat{\beta} - \beta)$ is asymptotically normal with bounded variance.

A deeper question is whether $\hat{\beta}$ achieves the lowest possible variance among all "reasonable" estimators. One important criteria for "reasonableness" is *regularity*, which is satisfied by $\hat{\beta}$ and most standard $\sqrt{n}$ consistent estimators and which we will define more precisely in Chapter 3. A regular estimator is *efficient* if it achieves the lowest possible variance among regular estimators. Semiparametric inference tools are required to establish this kind of optimality. Unfortunately, $\hat{\beta}$ does not have the lowest possible variance among all *regular* estimators, unless we are willing to make some very strong assumptions. For instance, $\hat{\beta}$ has optimal variance if we are willing to assume, in addition to what we have already assumed, that $e$ has a Gaussian distribution and is independent of $Z$. In this instance, the model is almost fully parametric (except that the distribution of $Z$ remains unspecified). Returning to the more general model in the previous paragraph, there is a modification of $\hat{\beta}$ that does have the lowest possible variance among regular estimators, but computation of this modified estimator requires estimation of the function $z \mapsto \mathrm{E}[e^2|Z = z]$. We will explore this particular example in greater detail in Chapters 3 and 4.

There is an interesting semiparametric model part way between the fully parametric Gaussian residual model and the more general model which only assumes $\mathrm{E}[e|Z] = 0$ almost surely. This alternative model assumes that the residual $e$ and covariate $Z$ are independent, with $\mathrm{E}[e] = 0$ and $\mathrm{E}[e^2] < \infty$, but no additional restrictions are placed on the residual distribution $F$. Unfortunately, $\hat{\beta}$ still does not have optimal variance. However, $\hat{\beta}$ is a very good estimator and may be good enough for most purposes. In this setting, it may be useful to estimate $F$ to determine whether the residuals are Gaussian. One promising estimator is

$$\hat{F}(t) = n^{-1} \sum_{i=1}^{n} 1\left\{Y_i - \hat{\beta}' Z_i \le t\right\}, \tag{1.2}$$

where $1\{A\}$ is the indicator of $A$. Empirical process methods can be used to show that $\hat{F}$ is uniformly consistent for $F$ and that $\sqrt{n}(\hat{F} - F)$ converges "in distribution" in a uniform sense to a certain Gaussian quantity, provided $f$ is uniformly bounded. Quotes are used here because the convergence in question involves random real functions rather than Euclidean random variables. This kind of convergence is called *weak convergence* and is a generalization of convergence in distribution which will be defined more precisely in Chapter 2.

Now we will consider a more complex example. Let $N(t)$ be a counting process over the finite interval $[0, \tau]$ which is free to jump as long as $t \in [0, V]$, where $V \in (0, \tau]$ is a random time. A counting process, by definition, is nonnegative and piecewise constant with positive jumps of size 1. Typically, the process counts a certain kind of event, such as hospitalizations, for an individual (see Chapter 1 of Fleming and Harrington, 1991, or Chapter II of Andersen, Borgan, Gill and Keiding, 1993). Define also the "at-risk" process $Y(t) = 1\{V \geq t\}$. This process indicates whether an individual is at-risk at time $t-$ (just to the left of $t$) for a jump in $N$ at time $t$. Suppose we also have baseline covariates $Z \in \mathbb{R}^p$, and, for all $t \in [0, \tau]$, we assume

$$\mathrm{E}\{N(t)|\, Z\} = \int_0^t \mathrm{E}\{Y(s)|\, Z\}\, e^{\beta' Z} d\Lambda(s), \tag{1.3}$$

for some $\beta \in \mathbb{R}^p$ and continuous nondecreasing function $\Lambda(t)$ with $\Lambda(0) = 0$ and $0 < \Lambda(\tau) < \infty$. The model (1.3) is a variant of the "multiplicative intensity model" (see definition 4.2.1 of Fleming and Harrington, 1991). Basically, we are assuming that the mean of the counting process is proportional to $e^{\beta' Z}$. We also need to assume that $\mathrm{E}\{Y(\tau)\} > 0$, $\mathrm{E}\{N^2(\tau)\} < \infty$, and that $Z$ is restricted to a bounded set, but we do not otherwise restrict the distribution of $N$. Given an i.i.d. sample $(N_i, Y_i, Z_i), i = 1, \ldots, n$, we are interested in estimating $\beta$ and $\Lambda$.

Under mild regularity conditions, the estimating equation

$$U_n(t, \beta) \;=\; n^{-1} \sum_{i=1}^{n} \int_0^t [Z_i - E_n(s, \beta)]\, dN_i(s), \tag{1.4}$$

where

$$E_n(s, \beta) = \frac{n^{-1} \sum_{i=1}^n Z_i Y_i(s) e^{\beta' Z_i}}{n^{-1} \sum_{i=1}^n Y_i(s) e^{\beta' Z_i}},$$

can be used for estimating $\beta$. The motivation for this estimating equation is that it arises as the score equation from the celebrated Cox partial likelihood (Cox, 1975) for either failure time data (where the counting process $N(t)$ simply indicates whether the failure time has occurred by time $t$) or the multiplicative intensity model under an independent increment assumption on $N$ (See Chapter 4 of Fleming and Harrington, 1991). Interestingly, this estimating equation can be shown to work under the more general model (1.3). Specifically, we can establish that (1.4) has an asymptotically unique zero $\hat{\beta}$ at $t = \tau$, that $\hat{\beta}$ is consistent for $\beta$, and that $\sqrt{n}\left(\hat{\beta} - \beta\right)$ is asymptotically mean zero Gaussian. Empirical process tools are needed to accomplish this. These same techniques can also establish that

$$\hat{\Lambda}(t) = \int_0^t \frac{n^{-1} \sum_{i=1}^n dN_i(s)}{n^{-1} \sum_{i=1}^n Y_i(s) e^{\hat{\beta}' Z_i}}$$

is uniformly consistent for $\Lambda(t)$ over $t \in [0, \tau]$ and that $\sqrt{n}(\hat{\Lambda} - \Lambda)$ converges weakly to a mean zero Gaussian quantity. These methods can also be used to construct valid confidence intervals and confidence bands for $\beta$ and $\Lambda$.

At this point, efficiency of these estimators is difficult to determine because so little has been specified about the distribution of $N$. Consider, however, the special case of right-censored failure time data. In this setting, the observed data are $(V_i, d_i, Z_i), i = 1, \ldots, n$, where $V_i = T_i \wedge C_i$, $d_i = 1\{T_i \leq C_i\}$, $T_i$ is a failure time of interest with integrated hazard function $e^{\beta' Z_i}\Lambda(t)$ given $Z_i$, and $C_i$ is a right censoring time independent of $T_i$ given $Z_i$ with distribution not depending on $\beta$ or $\Lambda$. Here, $N_i(t) = d_i 1\{V_i \leq t\}$ and $a \wedge b$ denotes the minimum of $a$ and $b$. Semiparametric inference techniques can now establish that both $\hat{\beta}$ and $\hat{\Lambda}$ are efficient. We will revisit this example in greater detail in Chapters 3 and 4.

In both of the previous examples, the estimator of the infinite-dimensional parameter ($F$ in the first example and $\Lambda$ in the second) is $\sqrt{n}$ consistent, but slower rates of convergence for the infinite dimensional part are also possible. As a third example, consider the partly linear logistic regression model described in Mammen and van de Geer (1997) and van der Vaart (1998, Page 405). The observed data are $n$ independent realizations of the random triplet $(Y, Z, U)$, where $Z \in \mathbb{R}^p$ and $U \in \mathbb{R}$ are covariates that are not linearly dependent, and $Y$ is a dichotomous outcome with

$$\mathrm{E}\{Y | Z, U\} = \nu[\beta' Z + \eta(U)], \tag{1.5}$$

where $\beta \in \mathbb{R}^p$, $Z$ is restricted to a bounded set, $U \in [0, 1]$, $\nu(t) = 1/(1+e^{-t})$, and $\eta$ is an unknown smooth function. We assume, for some integer $k \geq 1$, that the first $k-1$ derivatives of $\eta$ exist and are absolutely continuous with $J^2(\eta) \equiv \int_0^1 \left[\eta^{(k)}(t)\right]^2 dt < \infty$, where superscript $(k)$ denotes the $k$-th derivative. Given an i.i.d. sample $X_i = (Y_i, Z_i, U_i), i = 1 \ldots n$, we are interested in estimating $\beta$ and $\eta$.

The conditional density at $Y = y$ given the covariates $(Z, U) = (z, u)$ has the form

$$p_{\beta,\eta}(x) = \{\nu[\beta' z + \eta(u)]\}^y \{1 - \nu[\beta' z + \eta(u)]\}^{1-y}.$$

This cannot be used directly for defining a likelihood since for any $1 \leq n < \infty$ and fixed sample $x_1, \ldots, x_n$, there exists a sequence of smooth functions $\{\hat{\eta}_m\}$ satisfying our criteria which converges to $\hat{\eta}$, where $\hat{\eta}(u_i) = \infty$ when $y_i = 1$ and $\hat{\eta}(u_i) = -\infty$ when $y_i = 0$. The issue is that requiring $J(\hat{\eta}) < \infty$ does not restrict $\hat{\eta}$ on any finite collection of points. There are a number of methods for addressing this problem, including requiring $J(\hat{\eta}) \leq M_n$ for each $n$, where $M_n \uparrow \infty$ at an appropriately slow rate, or using a series of increasingly complex spline approximations.

An important alternative is to use the penalized log-likelihood

$$\tilde{L}_n(\beta, \eta) = n^{-1} \sum_{i=1}^{n} \log p_{\beta,\eta}(X_i) - \hat{\lambda}_n^2 J^2(\eta), \tag{1.6}$$

where $\hat{\lambda}_n$ is a possibly data-dependent *smoothing parameter*. $\tilde{L}_n$ is maximized over $\beta$ and $\eta$ to obtain the estimators $\hat{\beta}$ and $\hat{\eta}$. Large values of $\hat{\lambda}_n$ lead to very smooth but somewhat biased $\hat{\eta}$, while small values of $\hat{\lambda}_n$ lead to less smooth but less biased $\hat{\eta}$. The proper trade-off between smoothness and bias is usually best achieved by data-dependent schemes such as cross-validation. If $\hat{\lambda}_n$ is chosen to satisfy $\hat{\lambda}_n = o_P(n^{-1/4})$ and $\hat{\lambda}_n^{-1} = O_P(n^{k/(2k+1)})$, then both $\hat{\beta}$ and $\hat{\eta}$ are uniformly consistent and $\sqrt{n}\left(\hat{\beta} - \beta\right)$ converges to a mean zero Gaussian vector. Furthermore, $\hat{\beta}$ can be shown to be efficient even though $\hat{\eta}$ is not $\sqrt{n}$ consistent. More about this example will be discussed in Chapter 4.

These three examples illustrate the goals of empirical process and semiparametric inference research as well as hint at the power of these methods for solving statistical inference problems involving infinite-dimensional parameters. The goal of the first part of this book is to present the key ideas of empirical processes and semiparametric inference in a concise and heuristic way, without being distracted by technical issues, and to provide motivation to pursue the subject in greater depth as given in the remaining parts of the book. Even for those anxious to pursue the subject in depth, the broad view contained in this first part provides a valuable context for learning the details.

Chapter 2 presents an overview of empirical process methods and results, while Chapter 3 presents an overview of semiparametric inference techniques. Several case studies illustrating these methods, including further details on the examples given above, are presented in Chapter 4, which concludes the overview part. The empirical process part of the book (Part II) will be introduced in Chapter 5, while the semiparametric inference part (Part III) will be introduced in Chapter 16.

# 2
# An Overview of Empirical Processes

This chapter presents an overview of the main ideas and techniques of empirical process research. The emphasis is on those concepts which directly impact statistical estimation and inference. The major distinction between empirical process theory and more standard asymptotics is that the random quantities studied have realizations as functions rather than real numbers or vectors. Proofs of results and certain details in definitions are postponed until Part II of the book.

We begin by defining and sketching the main features and asymptotic concepts of empirical processes with a view towards statistical issues. An outline of the main empirical process techniques covered in this book is presented next. This chapter concludes with a discussion of several additional related topics that will not be pursued in later chapters.

## 2.1 The Main Features

A *stochastic process* is a collection of random variables $\{X(t), t \in T\}$ on the same probability space, indexed by an arbitrary index set $T$. An *empirical process* is a stochastic process based on a random sample. For example, consider a random sample $X_1, \ldots, X_n$ of i.i.d. real random variables with distribution $F$. The *empirical distribution function* is

$$\mathbb{F}_n(t) \quad = \quad n^{-1} \sum_{i=1}^{n} 1\{X_i \leq t\}, \tag{2.1}$$

where the index $t$ is allowed to vary over $T = \mathbb{R}$, the real line.

More generally, we can consider a random sample $X_1, \ldots, X_n$ of independent draws from a probability measure $P$ on an arbitrary sample space $\mathcal{X}$. We define the *empirical measure* to be $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$, where $\delta_x$ is the measure that assigns mass 1 at $x$ and zero elsewhere. For a measurable function $f : \mathcal{X} \mapsto \mathbb{R}$, we denote $\mathbb{P}_n f = n^{-1} \sum_{i=1}^n f(X_i)$. For any class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$, an empirical process $\{\mathbb{P}_n f, f \in \mathcal{F}\}$ can be defined. This simple approach can generate a surprising variety of empirical processes, many of which we will consider in later sections in this chapter as well as in Part II.

Setting $\mathcal{X} = \mathbb{R}$, we can now re-express $\mathbb{F}_n$ as the empirical process $\{\mathbb{P}_n f, f \in \mathcal{F}\}$, where $\mathcal{F} = \{1\{x \leq t\}, t \in \mathbb{R}\}$. Thus one can view the stochastic process $\mathbb{F}_n$ as indexed by either $t \in \mathbb{R}$ or $f \in \mathcal{F}$. We will use either indexing approach, depending on which is most convenient for the task at hand. However, because of its generality, indexing empirical processes by classes of functions will be the primary approach taken throughout this book.

By the law of large numbers, we know that

$$\mathbb{F}_n(t) \quad \stackrel{\text{as}}{\to} \quad F(t) \tag{2.2}$$

for each $t \in \mathbb{R}$, where $\stackrel{\text{as}}{\to}$ denotes almost sure convergence. A primary goal of empirical process research is to study empirical processes as random functions over the associated index set. Each realization of one of these random functions is a *sample path*. To this end, Glivenko (1933) and Cantelli (1933) demonstrated that (2.2) could be strengthened to

$$\sup_{t \in \mathbb{R}} |\mathbb{F}_n(t) - F(t)| \quad \stackrel{\text{as}}{\to} \quad 0. \tag{2.3}$$

Another way of saying this is that the sample paths of $F_n$ get uniformly closer to $F$ as $n \to \infty$. Returning to general empirical processes, a class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$, is said to be a *P-Glivenko-Cantelli* class if

$$\sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \quad \stackrel{\text{as*}}{\to} \quad 0, \tag{2.4}$$

where $Pf = \int_{\mathcal{X}} f(x) P(dx)$ and $\stackrel{\text{as*}}{\to}$ is a mode of convergence slightly stronger than $\stackrel{\text{as}}{\to}$ but which will not be precisely defined until later in this chapter (both modes of convergence are equivalent in the setting of (2.3)). Sometimes the $P$ in $P$-Glivenko-Cantelli can be dropped if the context is clear.

Returning to $\mathbb{F}_n$, we know by the central limit theorem that for each $t \in \mathbb{R}$

$$G_n(t) \equiv \sqrt{n}\,[\mathbb{F}_n(t) - F(t)] \quad \rightsquigarrow \quad G(t),$$

where $\rightsquigarrow$ denotes convergence in distribution and $G(t)$ is a mean zero normal random variable with variance $F(t)\,[1 - F(t)]$. In fact, we know that $G_n$, simultaneously for all $t$ in a finite set $T_k = \{t_1, \ldots, t_k\} \in \mathbb{R}$, will converge in distribution to a mean zero multivariate normal vector $G = \{G(t_1), \ldots, G(t_k)\}'$, where

$$\text{(2.5)} \qquad \operatorname{cov}[G(s), G(t)] = \mathrm{E}\,[G(s)G(t)] \quad = \quad F(s \wedge t) - F(s)F(t)$$

for all $s, t \in T_k$.

Much more can be said. Donsker (1952) showed that the sample paths of $G_n$, as functions on $\mathbb{R}$, converge in distribution to a certain stochastic process $G$. *Weak convergence* is the generalization of convergence in distribution from vectors of random variables to sample paths of stochastic processes. Donsker's result can be stated succinctly as $G_n \rightsquigarrow G$ in $\ell^\infty(\mathbb{R})$, where, for any index set $T$, $\ell^\infty(T)$ is the collection of all bounded functions $f : T \mapsto \mathbb{R}$. $\ell^\infty(T)$ is used in settings like this to remind us that we are thinking of distributional convergence in terms of the sample paths.

The limiting process $G$ is a mean zero *Gaussian process* with $\mathrm{E}\,[G(s)G(t)]$ = (2.5) for every $s, t \in \mathbb{R}$. A Gaussian process is a stochastic process $\{Z(t), t \in T\}$, where for every finite $T_k \subset T$, $\{Z(t), t \in T_k\}$ is multivariate normal, and where all sample paths are continuous in a certain sense that will be made more explicit later in this chapter. The process $G$ can be written $G(t) = \mathbb{B}(F(t))$, where $\mathbb{B}$ is a standard Brownian bridge on the unit interval. The process $\mathbb{B}$ has covariance $s \wedge t - st$ and is equivalent to the process $\mathbb{W}(t) - t\mathbb{W}(1)$, for $t \in [0, 1]$, where $\mathbb{W}$ is a *standard Brownian motion* process. The standard Brownian motion is a Guassian process on $[0, \infty)$ with continuous sample paths, with $\mathbb{W}(0) = 0$, and with covariance $s \wedge t$. Both $\mathbb{B}$ and $\mathbb{W}$ are important examples of Gaussian processes.

Returning again to general empirical processes, define the random measure $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$, and, for any class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$, let $\mathbb{G}$ be a mean zero Gaussian process indexed by $\mathcal{F}$, with covariance $\mathrm{E}\,[f(X)g(X)] - \mathrm{E}f(X)\mathrm{E}g(X)$ for all $f, g \in \mathcal{F}$, and having appropriately continuous sample paths. Both $\mathbb{G}_n$ and $\mathbb{G}$ can be thought of as being indexed by $\mathcal{F}$. We say that $\mathcal{F}$ is $P$-Donsker if $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathcal{F})$. The $P$ and/or the $\ell^\infty(\mathcal{F})$ may be dropped if the context is clear. Donsker's (1952) theorem tells us that $\mathcal{F} = \{1\{x \le t\}, t \in \mathbb{R}\}$ is Donsker for all probability measures which are based on some real distribution function $F$. With $f(x) = 1\{x \le t\}$ and $g(x) = 1\{x \le s\}$,

$$\mathrm{E}\,[f(X)g(X)] - \mathrm{E}f(X)\mathrm{E}g(X) = F(s \wedge t) - F(s)F(t).$$

For this reason, $\mathbb{G}$ is also referred to as a Brownian bridge.

Suppose we are interested in forming confidence bands for $F$ over some subset $H \subset \mathbb{R}$. Because $\mathcal{F} = \{1\{x \le t\}, t \in \mathbb{R}\}$ is Glivenko-Cantelli, we can uniformly consistently estimate the covariance $\sigma(s, t) = F(s \wedge t) - F(s)F(t)$ of $G$ with $\hat{\sigma}(s, t) = \mathbb{F}_n(s \wedge t) - \mathbb{F}_n(s)\mathbb{F}_n(t)$. While such a covariance could be

used to form confidence bands when $H$ is finite, it is of little use when $H$ is infinite, such as when $H$ is a subinterval of $\mathbb{R}$. In this case, it is preferable to make use of the Donsker result for $G_n$. Let $U_n = \sup_{t\in H} |G_n(t)|$. The *continuous mapping theorem* tells us that whenever a process $\{Z_n(t), t \in H\}$ converges weakly to a tight limiting process $\{Z(t), t \in H\}$ in $\ell^\infty(H)$, then $h(Z_n) \rightsquigarrow h(Z)$ in $h(\ell^\infty(H))$ for any continuous map $h$. In our setting $U_n = h(G_n)$, where $h(g) = \sup_{t\in H} |g(t)|$, for any $g \in \ell^\infty(\mathbb{R})$, is a continuous real function. Thus the continuous mapping theorem tells us that $U_n \rightsquigarrow U = \sup_{t\in H} |G(t)|$. When $F$ is continuous and $H = \mathbb{R}$, $U = \sup_{t\in[0,1]} |\mathbb{B}(t)|$ has a known distribution from which it is easy to compute quantiles. If we let $u_p$ be the $p$-th quantile of $U$, then an asymptotically valid symmetric $1-\alpha$ level confidence band for $F$ is $\mathbb{F}_n \pm u_{1-\alpha}/\sqrt{n}$.

An alternative is to construct confidence bands based on a large number of bootstraps of $\mathbb{F}_n$. The bootstrap for $\mathbb{F}_n$ can be written as $\hat{\mathbb{F}}_n(t) = n^{-1}\sum_{i=1}^n W_{ni} 1\{X_i \le t\}$, where $(W_{n1}, \ldots, W_{nn})$ is a multinomial random $n$-vector, with probabilities $1/n, \ldots, 1/n$ and number of trials $n$, and which is independent of the data $X_1, \ldots, X_n$. The conditional distribution of $\hat{G}_n = \sqrt{n}(\hat{\mathbb{F}}_n - \mathbb{F}_n)$ given $X_1, \ldots, X_n$ can be shown to converge weakly to the distribution of $G$ in $\ell^\infty(\mathbb{R})$. Thus the bootstrap is an asymptotically valid way to obtain confidence bands for $F$.

Returning to the general empirical process set-up, let $\mathcal{F}$ be a Donsker class and suppose we wish to construct confidence bands for $\mathrm{E}f(X)$ that are simultaneously valid for all $f \in \mathcal{H} \subset \mathcal{F}$. Provided certain second moment conditions hold on $\mathcal{F}$, the estimator $\hat{\sigma}(f,g) = \mathbb{P}_n[f(X)g(X)] - \mathbb{P}_n f(X)\mathbb{P}_n g(X)$ is consistent for $\sigma(f,g) = \mathrm{E}[f(X)g(X)] - \mathrm{E}f(X)\mathrm{E}g(X)$ uniformly over all $f, g \in \mathcal{F}$. As with the empirical distribution function estimator, this covariance is enough to form confidence bands provided $\mathcal{H}$ is finite. Fortunately, the bootstrap is always asymptotically valid when $\mathcal{F}$ is Donsker and can therefore be used for infinite $\mathcal{H}$. More precisely, if $\hat{\mathbb{G}}_n = \sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n)$, where $\hat{\mathbb{P}}_n f = n^{-1}\sum_{i=1}^n W_{ni} f(X_i)$ and $(W_{n1}, \ldots, W_{nn})$ is defined as before, then the conditional distribution of $\hat{\mathbb{G}}_n$ given the data converges weakly to $\mathbb{G}$ in $\ell^\infty(\mathcal{F})$. Since this is true for all of $\mathcal{F}$, it is certainly true for any $\mathcal{H} \subset \mathcal{F}$. The bootstrap result for $\mathbb{F}_n$ is clearly a special case of this more general result.

Many important statistics based on i.i.d. data cannot be written as empirical processes, but they can frequently be written in the form $\phi(\mathbb{P}_n)$, where $\mathbb{P}_n$ is indexed by some $\mathcal{F}$ and $\phi$ is a smooth map from $\ell^\infty(\mathcal{F})$ to some set $B$ (possibly infinite-dimensional). Consider, for example, the quantile process $\xi_n(p) = \mathbb{F}_n^{-1}(p)$ for $p \in [a,b]$, where $H^{-1}(p) = \inf\{t : H(t) \ge p\}$ for a distribution function $H$ and $0 < a < b < 1$. Here, $\xi_n = \phi(\mathbb{F}_n)$, where $\phi$ maps a distribution function $H$ to $H^{-1}$. When the underlying distribution $F$ is continuous over $N = [H^{-1}(a) - \epsilon, H^{-1}(b) + \epsilon] \subset [0,1]$, for some $\epsilon > 0$, with continuous density $f$ such that $0 < \inf_{t\in N} f(t) \le \sup_{t\in N} f(t) < \infty$, then $\sqrt{n}(\xi_n(p) - \xi_p)$, where $\xi_p = F^{-1}(p)$, is uniformly

asymptotically equivalent to $-G_n(F^{-1}(p))/f(F^{-1}(p))$ and hence converges weakly to $G(F^{-1}(p))/f(F^{-1}(p))$ in $\ell^\infty([a,b])$. (Because the process $G$ is symmetric around zero, both $-G$ and $G$ have the same distribution.) The above weak convergence result is a special case of the *functional delta-method* principle which states that $\sqrt{n}\,[\phi(\mathbb{P}_n) - \phi(P)]$ converges weakly in $B$ to $\phi'(\mathbb{G})$, whenever $\mathcal{F}$ is Donsker and $\phi$ has a "Hadamard derivative" $\phi'$ which will be defined more precisely later in this chapter.

Many additional statistics can be written as zeros or maximizers of certain data-dependent processes. The former are known as *Z-estimators* and the latter as *M-estimators*. Consider the linear regression example given in Chapter 1. Since $\hat{\beta}$ is the zero of $U_n(\beta) = \mathbb{P}_n\,[X(Y - X'\beta)]$, $\hat{\beta}$ is a Z-estimator. In contrast, the penalized likelihood estimators $(\hat{\beta}, \hat{\eta})$ in the partly linear logistic regression example of the same chapter are M-estimators since they are maximizers of $\tilde{L}(\beta, \eta)$ given in (1.6). As is the case with $U_n$ and $\tilde{L}_n$, the data-dependent objective functions used in Z- and M- estimation are often empirical processes, and thus empirical process methods are frequently required when studying the large sample properties of the associated statistics.

The key attribute of empirical processes is that they are random functions—or stochastic processes—based on a random data sample. The main asymptotic issue is studying the limiting behavior of these processes in terms of their sample paths. Primary achievements in this direction are Glivenko-Cantelli results which extend the law of large numbers, Donsker results which extend the central limit theorem, the validity of the bootstrap for Donsker classes, and the functional delta method.

## 2.2 Empirical Process Techniques

In this section, we expand on several important techniques used in empirical processes. We first define and discuss several important kinds of stochastic convergence, including convergence in probability as well as almost sure and weak convergence. We then introduce the concept of entropy and introduce several Glivenko-Cantelli and Donsker theorems based on entropy. The empirical bootstrap and functional delta method are described next. A brief outline of Z- and M- estimator methods are then presented. This section is essentially a review in miniature of the main points covered in Part II of this book, with a minimum of technicalities.

### *2.2.1 Stochastic Convergence*

When discussing convergence of stochastic processes, there is always a *metric space* $(\mathbb{D}, d)$ implicitly involved, where $\mathbb{D}$ is the space of possible values for the processes and $d$ is a *metric* (distance measure), satisfying $d(x, y) \geq 0$,

$d(x,y) = d(y,x)$, $d(x,z) \leq d(x,y) + d(y,z)$, and $d(x,y) = 0$ if and only if $x = y$, for all $x, y, z \in \mathbb{D}$. Frequently, $\mathbb{D} = \ell^\infty(T)$, where $T$ is the index set for the processes involved, and $d$ is the uniform distance on $\mathbb{D}$, i.e., $d(x,y) = \sup_{t\in T} |x(t) - y(t)|$ for any $x, y \in \mathbb{D}$. We are primarily interested in the convergence properties of the sample paths of stochastic processes. Weak convergence, or convergence in distribution, of a stochastic process $X_n$ happens when the sample paths of $X_n$ begin to behave in distribution, as $n \to \infty$, more and more like a specific random process $X$. When $X_n$ and $X$ are *Borel measurable*, weak convergence is equivalent to saying that $\mathrm{E}f(X_n) \to \mathrm{E}f(X)$ for every bounded, continuous function $f : \mathbb{D} \mapsto \mathbb{R}$, where the notation $f : A \mapsto B$ means that $f$ is a mapping from $A$ to $B$, and where continuity is in terms of $d$. Hereafter, we will let $C_b(\mathbb{D})$ denote the space of bounded, continuous maps $f : \mathbb{D} \mapsto \mathbb{R}$. We will define Borel measurability in detail later in Part II, but, for now, it is enough to say that lack of this property means that there are certain important subsets $A \subset \mathbb{D}$ where the probability that $X_n \in A$ is not defined.

In many statistical applications, $X_n$ may not be Borel measurable. To resolve this problem, we need to introduce the notion of *outer expectation* for arbitrary maps $T : \Omega \mapsto \bar{\mathbb{R}} \equiv [-\infty, \infty]$, where $\Omega$ is the sample space. $T$ is not necessarily a random variable because it is not necessarily Borel measurable. The outer expectation of $T$, denoted $\mathrm{E}^*T$, is the infimum over all $\mathrm{E}U$, where $U : \Omega \mapsto \mathbb{R}$ is measurable, $U \geq T$, and $\mathrm{E}U$ exists. For $\mathrm{E}U$ to exist, it must not be indeterminate, although it can be $\pm\infty$, provided the sign is clear. We analogously define inner expectation: $\mathrm{E}_*T = -\mathrm{E}^*[-T]$. There also exists a measurable function $T^* : \Omega \mapsto \mathbb{R}$, called the *minimal measurable majorant*, satisfying $T^*(\omega) \geq T(\omega)$ for all $\omega \in \Omega$ and which is almost surely the smallest measurable function $\geq T$. Furthermore, when $\mathrm{E}^*T < \infty$, $\mathrm{E}^*T = \mathrm{E}T^*$. The *maximal measurable minorant* is $T_* = -(-T)^*$. We also define outer probability for possibly nonmeasurable sets: $\mathrm{P}^*(A)$ as the infimum over all $\mathrm{P}(B)$ with $A \subset B \subset \Omega$ and $B$ a Borel measurable set. Inner probability is defined as $\mathrm{P}_*(A) = 1 - \mathrm{P}^*(\Omega - A)$. This use of outer measure permits defining weak convergence, for possibly nonmeasurable $X_n$, as $\mathrm{E}^*f(X_n) \mapsto \mathrm{E}f(X)$ for all $f \in C_b(\mathbb{D})$. We denote this convergence by $X_n \rightsquigarrow X$. Notice that we require the limiting process $X$ to be measurable. This definition of weak convergence also carries with it an implicit measurability requirement on $X_n$: $X_n \rightsquigarrow X$ implies that $X_n$ is *asymptotically measurable*, in that $\mathrm{E}^*f(X_n) - \mathrm{E}_*f(X_n) \to 0$, for every $f \in C_b(\mathbb{D})$.

We now consider convergence in probability and almost surely. We say $X_n$ converges to $X$ in probability if $\mathrm{P}\{d(X_n, X)^* > \epsilon\} \to 0$ for every $\epsilon > 0$, and we denote this $X_n \xrightarrow{\mathrm{P}} X$. We say that $X_n$ converges outer almost surely to $X$ if there exists a sequence $\Delta_n$ of measurable random variables with $d(X_n, X) \leq \Delta_n$ for all $n$ and with $\mathrm{P}\{\limsup_{n\to\infty} \Delta_n = 0\} = 1$. We denote this kind of convergence $X_n \xrightarrow{\mathrm{as}^*} X$. While these modes of convergence are

slightly different than the standard ones, they are identical when all the quantities involved are measurable. The properties of the standard modes are also generally preserved in these new modes. The major difference is that these new modes can accommodate many situations in statistics and in other fields which could not be as easily accommodated with the standard ones. As much as possible, we will keep measurability issues suppressed throughout this book, except where it is necessary for clarity. From this point on, the metric $d$ of choice will be the uniform metric unless noted otherwise.

For almost all of the weak convergence applications in this book, the limiting quantity $X$ will be *tight*, in the sense that the sample paths of $X$ will have a certain minimum amount of smoothness. To be more precise, for an index set $T$, let $\rho$ be a *semimetric* on $T$, in that $\rho$ has all the properties of a metric except that $\rho(s,t) = 0$ does not necessarily imply $s = t$. We say that $T$ is totally bounded by $\rho$ if for every $\epsilon > 0$, there exists a finite collection $T_k = \{t_1, \ldots, t_k\} \subset T$ such that for all $t \in T$, we have $\rho(t,s) \leq \epsilon$ for some $s \in T_k$. Now define $UC(T,\rho)$ to be the subset of $\ell^\infty(T)$ where each $x \in UC(T,\rho)$ satisfies

$$\lim_{\delta \downarrow 0} \sup_{s,t \in T \text{ with } \rho(s,t) \leq \delta} |x(t) - x(s)| = 0.$$

The "$UC$" refers to uniform continuity. The stochastic process $X$ is tight if $X \in UC(T,\rho)$ almost surely for some $\rho$ for which $T$ is totally bounded. If $X$ is a Gaussian process, then $\rho$ can be chosen as $\rho(s,t) = (\text{var}[X(s) - X(t)])^{1/2}$. Tight Gaussian processes will be the most important limiting processes considered in this book.

Two conditions need to be met in order for $X_n$ to converge weakly in $\ell^\infty(T)$ to a tight $X$. This is summarized in the following theorem which we present now but prove later in Chapter 7 (Page 114):

THEOREM 2.1 *$X_n$ converges weakly to a tight $X$ in $\ell^\infty(T)$ if and only if:*

*(i) For all finite $\{t_1, \ldots, t_k\} \subset T$, the multivariate distribution of $\{X_n(t_1), \ldots, X_n(t_k)\}$ converges to that of $\{X(t_1), \ldots, X(t_k)\}$.*

*(ii) There exists a semimetric $\rho$ for which $T$ is totally bounded and*

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} \mathrm{P}^* \left\{ \sup_{s,t \in T \text{ with } \rho(s,t) < \delta} |X_n(s) - X_n(t)| > \epsilon \right\} = 0, \tag{2.6}$$

*for all $\epsilon > 0$.*

Condition (i) is convergence of all finite dimensional distributions and Condition (ii) implies *asymptotic tightness*. In many applications, Condition (i) is not hard to verify while Condition (ii) is much more difficult.

In the empirical process setting based on i.i.d. data, we are interested in establishing that $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathcal{F})$, where $\mathcal{F}$ is some class of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$, and where $\mathcal{X}$ is the sample space. When $\mathrm{E}f^2(X) < \infty$ for all $f \in \mathcal{F}$, Condition (i) above is automatically satisfied by the standard central limit theorem, whereas establishing Condition (ii) is much more work and is the primary motivator behind the development of much of modern empirical process theory. Whenever $\mathcal{F}$ is Donsker, the limiting process $\mathbb{G}$ is always a tight Gaussian process, and $\mathcal{F}$ is totally bounded by the semimetric $\rho(f,g) = \{\mathrm{var}\,[f(X) - g(X)]\}^{1/2}$. Thus Conditions (i) and (ii) of Theorem 2.1 are both satisfied with $T = \mathcal{F}$, $X_n(f) = \mathbb{G}_n f$, and $X(f) = \mathbb{G}f$, for all $f \in \mathcal{F}$.

Another important result is the *continuous mapping theorem*. This theorem states that if $g : \mathbb{D} \mapsto \mathbb{E}$ is continuous at every point of a set $\mathbb{D}_0 \subset \mathbb{D}$, and if $X_n \rightsquigarrow X$, where $X$ takes all its values in $\mathbb{D}_0$, then $g(X_n) \rightsquigarrow g(X)$. For example, if $\mathcal{F}$ is a Donsker class, then $\sup_{f\in\mathcal{F}} |\mathbb{G}_n f|$ has the same limiting distribution as $\sup_{f\in\mathcal{F}} |\mathbb{G}f|$, since the supremum map is uniformly continuous, i.e., $\left|\sup_{f\in\mathcal{F}} |x(f)| - \sup_{f\in\mathcal{F}} |y(f)|\right| \leq \sup_{f\in\mathcal{F}} |x(f) - y(f)|$ for all $x, y \in \ell^\infty(\mathcal{F})$. This fact can be used to construct confidence bands for $Pf$. The continuous mapping theorem has many other practical uses that we will utilize at various points throughout this book.

### *2.2.2 Entropy for Glivenko-Cantelli and Donsker Theorems*

The major challenge in obtaining Glivenko-Cantelli or Donsker theorems for classes of functions $\mathcal{F}$ is to somehow show that going from pointwise convergence to uniform convergence is feasible. Clearly the complexity, or *entropy*, of $\mathcal{F}$ plays a major role. The easiest entropy to introduce is *entropy with bracketing*. For $1 \leq r < \infty$, Let $L_r(P)$ denote the collection of functions $g : \mathcal{X} \mapsto \mathbb{R}$ such that $\|g\|_{r,P} \equiv \left[\int_{\mathcal{X}} |g(x)|^r dP(x)\right]^{1/r} < \infty$. An $\epsilon$-*bracket* in $L_r(P)$ is a pair of functions $l, u \in L_r(P)$ with $\mathrm{P}\{l(X) \leq u(X)\} = 1$ and with $\|l - u\|_{r,P} \leq \epsilon$. A function $f \in \mathcal{F}$ lies in the bracket $l, u$ if $\mathrm{P}\{l(X) \leq f(X) \leq u(X)\} = 1$. The *bracketing number* $N_{[]}(\epsilon, \mathcal{F}, L_r(P))$ is the minimum number of $\epsilon$-brackets in $L_r(P)$ needed to ensure that every $f \in \mathcal{F}$ lies in at least one bracket. The logarithm of the bracketing number is the entropy with bracketing. The following is one of the simplest Glivenko-Cantelli theorems (the proof is deferred until Part II, Page 145):

THEOREM 2.2 *Let $\mathcal{F}$ be a class of measurable functions and suppose that $N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\epsilon > 0$. Then $\mathcal{F}$ is P-Glivenko-Cantelli.*

Consider, for example, the empirical distribution function $\mathbb{F}_n$ based on an i.i.d. sample $X_1, \ldots, X_n$ of real random variables with distribution $F$ (which defines the probability measure $P$ on $\mathcal{X} = \mathbb{R}$). In this setting, $\mathbb{F}_n$ is the empirical process $\mathbb{G}_n$ with class $\mathcal{F} = \{1\{x \leq t\}, t \in \mathbb{R}\}$. For any $\epsilon > 0$, a finite collection of real numbers $-\infty = t_1 < t_2 < \cdots < t_k = \infty$

can be found so that $F(t_j-) - F(t_{j-1}) \leq \epsilon$ for all $1 < j \leq k$, $F(t_1) = 0$ and $F(t_k-) = 1$, where $H(t-) = \lim_{s \uparrow t} H(s)$ when such a limit exists. This can always be done in such a way that $k \leq 2 + 1/\epsilon$. Consider the collection of brackets $\{(l_j, u_j), 1 < j \leq k\}$, with $l_j(x) = 1\{x \leq t_{j-1}\}$ and $u_j(x) = 1\{x < t_j\}$ (notice that $u_j$ is not in $\mathcal{F}$). Now each $f \in \mathcal{F}$ is in at least one bracket and $\|u_j - l_j\|_{P,1} = F(t_j-) - F(t_{j-1}) \leq \epsilon$ for all $1 < j \leq k$. Thus $N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\epsilon > 0$, and the conditions of Theorem 2.2 are met.

Donsker theorems based on entropy with bracketing require more stringent conditions on the number of brackets needed to cover $\mathcal{F}$. The *bracketing integral*,

$$J_{[]}(\delta, \mathcal{F}, L_r(P)) \equiv \int_0^\delta \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L_r(P))} d\epsilon,$$

needs to be bounded for $r = 2$ and $\delta = \infty$ to establish that $\mathcal{F}$ is Donsker. Hence the bracketing entropy is permitted to go to $\infty$ as $\epsilon \downarrow 0$, but not too quickly. For most of the classes $\mathcal{F}$ of interest, the entropy does go to $\infty$ as $\epsilon \downarrow 0$. However, a surprisingly large number of these classes satisfy the conditions of Theorem 2.3 below, our first Donsker theorem (which we prove in Chapter 8, Page 148):

THEOREM 2.3 *Let $\mathcal{F}$ be a class of measurable functions with $J_{[]}(\infty, \mathcal{F}, L_2(P)) < \infty$. Then $\mathcal{F}$ is $P$-Donsker.*

Returning again to the empirical distribution function example, we have for the $\epsilon$-brackets used previously that $\|u_j - l_j\|_{P,2} = (\|u_j - l_j\|_{P,1})^{1/2} \leq \epsilon^{1/2}$. Hence the minimum number of $L_2$ $\epsilon$-brackets needed to cover $\mathcal{F}$ is bounded by $1 + 1/\epsilon^2$, since an $L_1$ $\epsilon^2$-bracket is an $L_2$ $\epsilon$-bracket. For $\epsilon > 1$, the number of brackets needed is just 1. $J_{[]}(\infty, \mathcal{F}, L_2(P))$ will therefore be finite if $\int_0^1 \sqrt{\log(1 + 1/\epsilon^2)} d\epsilon < \infty$. Using the fact that $\log(1+a) \leq 1 + \log(a)$ for $a \geq 1$ and the variable substitution $u = 1 + \log(1/\epsilon^2)$, we obtain that this integral is bounded by $\int_0^\infty u^{1/2} e^{-u/2} du = \sqrt{2\pi}$. Thus the conditions of Theorem 2.3 are easily satisfied. We now give two other examples of classes with bounded $L_r(P)$ bracketing integral. Parametric classes of the form $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ work, provided $\Theta$ is a bounded subset of $\mathbb{R}^p$ and there exists an $m \in L_r(P)$ such that $|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq m(x)\|\theta_1 - \theta_2\|$ for all $\theta_1, \theta_2 \in \Theta$. Here, $\|\cdot\|$ is the standard Euclidean norm on $\mathbb{R}^p$. The class $\mathcal{F}$ of all monotone functions $f : \mathbb{R} \mapsto [0, 1]$ also works for all $1 \leq r < \infty$ and all probability measures $P$.

Entropy calculations for other classes that arise in statistical applications can be difficult. However, there are a number of techniques for doing this that are not difficult to apply in practice and that we will explore briefly later on in this section. Unfortunately, there are also many classes $\mathcal{F}$ for which entropy with bracketing does not work at all. An alternative which can be useful in such settings is entropy based on *covering numbers*. For a

probability measure $Q$, the covering number $N(\epsilon, \mathcal{F}, L_r(Q))$ is the minimum number of $L_r(Q)$ $\epsilon$-balls needed to cover $\mathcal{F}$, where an $L_r(Q)$ $\epsilon$-ball around a function $g \in L_r(Q)$ is the set $\{h \in L_r(Q) : \|h-g\|_{Q,r} < \epsilon\}$. For a collection of balls to cover $\mathcal{F}$, all elements of $\mathcal{F}$ must be included in at least one of the balls, but it is not necessary that the centers of the balls be contained in $\mathcal{F}$. The *entropy* is the logarithm of the covering number. The bracketing entropy conditions in Theorems 2.2 and 2.3 can be replaced by conditions based on the *uniform covering numbers*

$$\sup_Q N\left(\epsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q)\right), \tag{2.7}$$

where $F : \mathcal{X} \mapsto \mathbb{R}$ is an *envelope* for $\mathcal{F}$, meaning that $|f(x)| \leq F(x)$ for all $x \in \mathcal{X}$ and all $f \in \mathcal{F}$, and where the supremum is taken over all finitely discrete probability measures $Q$ with $\|F\|_{Q,r} > 0$. A finitely discrete probability measure on $\mathcal{X}$ puts mass only at a finite number of points in $\mathcal{X}$. Notice that the uniform covering number does not depend on the probability measure $P$ for the observed data. The *uniform entropy integral* is

$$J(\delta, \mathcal{F}, L_r) = \int_0^\delta \sqrt{\log \sup_Q N\left(\epsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q)\right)} d\epsilon,$$

where the supremum is taken over the same set used in (2.7).

The following two theorems (given without proof) are Glivenko-Cantelli and Donsker results for uniform entropy:

THEOREM 2.4 *Let $\mathcal{F}$ be an appropriately measurable class of measurable functions with $\sup_Q N(\epsilon\|F\|_{1,Q}, \mathcal{F}, L_1(Q)) < \infty$ for every $\epsilon > 0$, where the supremum is taken over the same set used in (2.7). If $P^*F < \infty$, then $\mathcal{F}$ is $P$-Glivenko-Cantelli.*

THEOREM 2.5 *Let $\mathcal{F}$ be an appropriately measurable class of measurable functions with $J(1, \mathcal{F}, L_2) < \infty$. If $P^*F^2 < \infty$, then $\mathcal{F}$ is $P$-Donsker.*

Discussion of the "appropriately measurable" condition will be postponed until Part II (see Pages 145 and 149), but suffice it to say that it is satisfied for many function classes of interest in statistical applications.

An important collection of function classes $\mathcal{F}$, which satisfies $J(1, \mathcal{F}, L_r) < \infty$ for any $1 \leq r < \infty$, are the *Vapnik-Červonenkis* classes, or VC classes. Many classes of interest in statistics are VC, including the class of indicator functions explored earlier in the empirical distribution function example and also vector space classes. A vector space class $\mathcal{F}$ has the form $\{\sum_{i=1}^k \lambda_i f_i(x), (\lambda_1, \ldots, \lambda_k) \in \mathbb{R}^k\}$ for fixed functions $f_1, \ldots, f_k$. We will postpone further definition and discussion of VC classes until Part II.

The important thing to know at this point is that one does not need to calculate entropy for each new problem. There are a number of easy

methods which can be used to determine whether a given class is Glivenko-Cantelli or Donsker based on whether the class is built up of other, well-known classes. For example, subsets of Donsker classes are Donsker since Condition (ii) of Theorem 2.1 is clearly satisfied for any subset of $T$ if it is satisfied for $T$. One can also use Theorem 2.1 to show that finite unions of Donsker classes are Donsker. When $\mathcal{F}$ and $\mathcal{G}$ are Donsker, the following are also Donsker: $\{f \wedge g : f \in \mathcal{F}, g \in \mathcal{G}\}$, $\{f \vee g : f \in \mathcal{F}, g \in \mathcal{G}\}$, where $\vee$ denotes maximum, and $\{f + g : f \in \mathcal{F}, g \in \mathcal{G}\}$. If $\mathcal{F}$ and $\mathcal{G}$ are bounded Donsker classes, then $\{fg : f \in \mathcal{F}, g \in \mathcal{G}\}$ is Donsker. Also, Lipschitz continuous functions of Donsker classes are Donsker. Furthermore, if $\mathcal{F}$ is Donsker, then it is also Glivenko-Cantelli. These, and many other tools for verifying that a given class is Glivenko-Cantelli or Donsker, will be discussed in greater detail in Chapter 9.

### *2.2.3 Bootstrapping Empirical Processes*

An important aspect of inference for empirical processes is to be able to obtain covariance and confidence band estimates. The limiting covariance for a $P$-Donsker class $\mathcal{F}$ is $\sigma : \mathcal{F} \times \mathcal{F} \mapsto \mathbb{R}$, where $\sigma(f, g) \equiv Pfg - PfPg$. The covariance estimate $\hat{\sigma} : \mathcal{F} \times \mathcal{F} \mapsto \mathbb{R}$, where $\hat{\sigma}(f, g) \equiv \mathbb{P}_n fg - \mathbb{P}_n f \mathbb{P}_n g$, is uniformly consistent for $\sigma$ outer almost surely if and only if $P^* \left[\sup_{f \in \mathcal{F}} (f(X) - Pf)^2\right] < \infty$. This will be proved later in Part II. However, this is only of limited use since critical values for confidence bands cannot in general be determined from the covariance when $\mathcal{F}$ is not finite. The bootstrap is an effective alternative.

As mentioned earlier, some care must be taken to ensure that the concept of weak convergence makes sense when the statistics of interest may not be measurable. This issue becomes more delicate with bootstrap results which involve convergence of conditional laws given the observed data. In this setting, there are two sources of randomness, the observed data and the resampling done by the bootstrap. For this reason, convergence of conditional laws is assessed in a slightly different manner than regular weak convergence. An important result is that $X_n \rightsquigarrow X$ in the metric space $(\mathbb{D}, d)$ if and only if

$$\sup_{f \in BL_1} |\mathrm{E}^* f(X_n) - \mathrm{E} f(X)| \quad \to \quad 0, \tag{2.8}$$

where $BL_1$ is the space of functions $f : \mathbb{D} \mapsto \mathbb{R}$ with Lipschitz norm bounded by 1, i.e., $\|f\|_\infty \leq 1$ and $|f(x) - f(y)| \leq d(x, y)$ for all $x, y \in \mathbb{D}$, and where $\|\cdot\|_\infty$ is the uniform norm.

We can now use this alternative definition of weak convergence to define convergence of the conditional limit laws of bootstraps. Let $\hat{X}_n$ be a sequence of bootstrapped processes in $\mathbb{D}$ with random weights that we will denote $M$. For some tight process $X$ in $\mathbb{D}$, we use the notation $\hat{X}_n \overset{\mathrm{P}}{\underset{M}{\rightsquigarrow}} X$ to

mean that $\sup_{h\in BL_1}\left|\mathrm{E}_M h(\hat{X}_n) - \mathrm{E}h(X)\right| \xrightarrow{\mathrm{P}} 0$ and $\mathrm{E}_M h(\hat{X}_n)^* - \mathrm{E}_M h(\hat{X}_n)_* \xrightarrow{\mathrm{P}} 0$, for all $h \in BL_1$, where the subscript $M$ in the expectations indicates conditional expectation over the weights $M$ given the remaining data, and where $h(\hat{X}_n)^*$ and $h(\hat{X}_n)_*$ denote measurable majorants and minorants with respect to the joint data (including the weights $M$). We use the notation $\hat{X}_n \underset{M}{\overset{\mathrm{as}*}{\rightsquigarrow}} X$ to mean the same thing except with all $\xrightarrow{\mathrm{P}}$'s replaced by $\xrightarrow{\mathrm{as}*}$'s. Note that the $h(\hat{X}_n)$ inside of the supremum does not have an asterisk: this is because Lipschitz continuous function of the bootstrapped processes we will study in this book will always be measurable functions of the random weights when conditioning on the data.

As mentioned previously, the bootstrap empirical measure can be defined as $\hat{\mathbb{P}}_n f = n^{-1}\sum_{i=1}^n W_{ni} f(X_i)$, where $\vec{W}_n = (W_{n1}, \ldots, W_{nn})$ is a multinomial vector with probabilities $(1/n, \ldots, 1/n)$ and number of trials $n$, and where $\vec{W}_n$ is independent of the data sequence $\vec{X} = (X_1, X_2, \ldots)$. We can now define a useful and simple alternative to this standard nonparametric bootstrap. Let $\vec{\xi} = (\xi_1, \xi_2, \ldots)$ be an infinite sequence of nonnegative i.i.d. random variables, also independent of $\vec{X}$, which have mean $0 < \mu < \infty$ and variance $0 < \tau^2 < \infty$, and which satisfy $\|\xi\|_{2,1} < \infty$, where $\|\xi\|_{2,1} = \int_0^\infty \sqrt{P(|\xi| > x)}dx$. This last condition is slightly stronger than bounded second moment but is implied whenever the $2+\epsilon$ moment exists for any $\epsilon > 0$. We can now define a *multiplier bootstrap* empirical measure $\tilde{\mathbb{P}}_n f = n^{-1}\sum_{i=1}^n (\xi_i/\bar{\xi}_n) f(X_i)$, where $\bar{\xi}_n = n^{-1}\sum_{i=1}^n \xi_i$ and $\tilde{\mathbb{P}}_n$ is defined to be zero if $\bar{\xi} = 0$. Note that the weights add up to $n$ for both bootstraps. When $\xi_1$ has a standard exponential distribution, for example, the moment conditions are clearly satisfied, and the resulting multiplier bootstrap has Dirichlet weights.

Under these conditions, we have the following two theorems (which we prove in Part II, Page 187), for convergence of the bootstrap, both in probability and outer almost surely. Let $\hat{\mathbb{G}}_n = \sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n)$, $\tilde{\mathbb{G}}_n = \sqrt{n}(\mu/\tau)(\tilde{\mathbb{P}}_n - \mathbb{P}_n)$, and $\mathbb{G}$ be the standard Brownian bridge in $\ell^\infty(\mathcal{F})$.

THEOREM 2.6 *The following are equivalent:*

*(i)* $\mathcal{F}$ *is P-Donsker.*

*(ii)* $\hat{\mathbb{G}}_n \underset{W}{\overset{\mathrm{P}}{\rightsquigarrow}} \mathbb{G}$ *in* $\ell^\infty(\mathcal{F})$ *and the sequence* $\hat{\mathbb{G}}_n$ *is asymptotically measurable.*

*(iii)* $\tilde{\mathbb{G}}_n \underset{\xi}{\overset{\mathrm{P}}{\rightsquigarrow}} \mathbb{G}$ *in* $\ell^\infty(\mathcal{F})$ *and the sequence* $\tilde{\mathbb{G}}_n$ *is asymptotically measurable.*

THEOREM 2.7 *The following are equivalent:*

*(i)* $\mathcal{F}$ *is P-Donsker and* $P^*\left[\sup_{f\in\mathcal{F}}(f(X) - Pf)^2\right] < \infty$.

*(ii)* $\hat{\mathbb{G}}_n \underset{W}{\overset{\mathrm{as}*}{\rightsquigarrow}} \mathbb{G}$ *in* $\ell^\infty(\mathcal{F})$.

*(iii)* $\tilde{\mathbb{G}}_n \overset{\text{as}^*}{\underset{\xi}{\rightsquigarrow}} \mathbb{G}$ *in* $\ell^\infty(\mathcal{F})$.

According to Theorem 2.7, the almost sure consistency of the bootstrap requires the same moment condition required for almost sure uniform consistency of the covariance estimator $\hat{\sigma}$. In contrast, the consistency in probability of the bootstrap given in Theorem 2.6 only requires that $\mathcal{F}$ is Donsker. Thus consistency in probability of the bootstrap empirical process is an automatic consequence of weak convergence in the first place. Fortunately, consistency in probability is adequate for most statistical applications, since this implies that confidence bands constructed from the bootstrap are asymptotically valid. This follows because, as we will also establish in Part II, whenever the conditional law of a bootstrapped quantity (say $\hat{X}_n$) in a normed space (with norm $\|\cdot\|$) converges to a limiting law (say of $X$), either in probability or outer almost surely, then the conditional law of $\|\hat{X}_n\|$ converges to that of $\|X\|$ under mild regularity conditions. We will also establish a slightly more general in-probability continuous mapping theorem for the bootstrap when the continuous map $g$ is real valued.

Suppose we wish to construct a $1-\alpha$ level confidence band for $\{Pf, f \in \mathcal{F}\}$, where $\mathcal{F}$ is $P$-Donsker. We can obtain a large number, say $N$, bootstrap realizations of $\sup_{f\in\mathcal{F}} \left|\hat{\mathbb{G}}_n f\right|$ to estimate the $1-\alpha$ quantile of $\sup_{f\in\mathcal{F}} |\mathbb{G}f|$. If we call this estimate $\hat{c}_{1-\alpha}$, then Theorem 2.6 tells us that $\{\mathbb{P}_n f \pm \hat{c}_{1-\alpha}, f \in \mathcal{F}\}$ has coverage $1-\alpha$ for large enough $n$ and $N$. For a more specific example, consider estimating $F(t_1,t_2) = \mathrm{P}\{Y_1 \le t_1, Y_2 \le t_2\}$, where $X = (Y_1, Y_2)$ has an arbitrary bivariate distribution. We can estimate $F(t_1,t_2)$ with $\hat{F}_n(t_1,t_2) = n^{-1}\sum_{i=1}^n 1\{Y_{1i} \le t_1, Y_{2i} \le t_2\}$. This is the same as estimating $\{Pf, f \in \mathcal{F}\}$, where $\mathcal{F} = \{f(x) = 1\{y_1 \le t_1, y_2 \le t_2\} : t_1, t_2 \in \mathbb{R}\}$. This is a bounded Donsker class since $\mathcal{F} = \{f_1 f_2 : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$, where $\mathcal{F}_j = \{1\{y_j \le t\}, t \in \mathbb{R}\}$ is a bounded Donsker class for $j = 1, 2$. We thus obtain consistency in probability of the bootstrap. We also obtain outer almost sure consistency of the bootstrap by Theorem 2.7, since $\mathcal{F}$ is bounded by 1.

### *2.2.4 The Functional Delta Method*

Suppose $X_n$ is a sequence of random variables with $\sqrt{n}(X_n - \theta) \rightsquigarrow X$ for some $\theta \in \mathbb{R}^p$, and the function $\phi : \mathbb{R}^p \mapsto \mathbb{R}^q$ has a derivative $\phi'(\theta)$ at $\theta$. The standard delta method now tells us that $\sqrt{n}(\phi(X_n) - \phi(\theta)) \rightsquigarrow \phi'(\theta)X$. However, many important statistics based on i.i.d. data involve maps from empirical processes to spaces of functions, and hence cannot be handled by the standard delta method. A simple example is the map $\phi_\xi$ which takes cumulative distribution functions $H$ and computes $\{\xi_p, p \in [a,b]\}$, where $\xi_p = H^{-1}(p) = \inf\{t : H(t) \ge p\}$ and $[a,b] \subset (0,1)$. The sample $p$-th quantile is then $\hat{\xi}_n(p) = \phi_\xi(\mathbb{F}_n)(p)$. Although the standard delta method cannot be used here, the functional delta method can be.

Before giving the main functional delta method results, we need to define derivatives for functions between normed spaces $\mathbb{D}$ and $\mathbb{E}$. A normed space is a metric space $(\mathbb{D}, d)$, where $d(x, y) = \|x-y\|$, for any $x, y \in \mathbb{D}$, and where $\|\cdot\|$ is a norm. A norm satisfies $\|x + y\| \leq \|x\| + \|y\|$, $\|\alpha x\| = |\alpha| \times \|x\|$, $\|x\| \geq 0$, and $\|x\| = 0$ if and only if $x = 0$, for all $x, y \in \mathbb{D}$ and all complex numbers $\alpha$. A map $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ is *Gâteaux-differentiable* at $\theta \in \mathbb{D}$, if for every fixed $h \in \mathbb{D}$ with $\theta + th \in \mathbb{D}_\phi$ for all $t > 0$ small enough, there exists an element $\phi'_\theta(h) \in \mathbb{E}$ such that

$$\frac{\phi(\theta + th) - \phi(\theta)}{t} \to \phi'_\theta(h)$$

as $t \downarrow 0$. For the functional delta method, however, we need $\phi$ to have the stronger property of being *Hadamard-differentiable*. A map $\phi : \mathbb{D}_\phi \mapsto \mathbb{E}$ is Hadamard-differentiable at $\theta \in \mathbb{D}$, tangentially to a set $\mathbb{D}_0 \subset \mathbb{D}$, if there exists a continuous linear map $\phi'_\theta : \mathbb{D} \mapsto \mathbb{E}$ such that

$$\frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} \to \phi'_\theta(h),$$

as $n \to \infty$, for all converging sequences $t_n \to 0$ and $h_n \to h \in \mathbb{D}_0$, with $h_n \in \mathbb{D}$ and $\theta + t_n h_n \in \mathbb{D}_\phi$ for all $n \geq 1$ sufficiently large.

For example, let $\mathbb{D} = D[0, 1]$, where $DA$, for any interval $A \subset \mathbb{R}$, is the space of cadlag (right-continuous with left-hand limits) real functions on $A$ with the uniform norm. Let $\mathbb{D}_\phi = \{f \in D[0, 1] : |f| > 0\}$. Consider the function $\phi : \mathbb{D}_\phi \mapsto \mathbb{E} = D[0, 1]$ defined by $\phi(g) = 1/g$. Notice that for any $\theta \in \mathbb{D}_\phi$, we have, for any converging sequences $t_n \downarrow 0$ and $h_n \to h \in \mathbb{D}$, with $h_n \in \mathbb{D}$ and $\theta + t_n h_n \in \mathbb{D}_\phi$ for all $n \geq 1$,

$$\frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} = \frac{1}{t_n(\theta + t_n h_n)} - \frac{1}{t_n \theta} = -\frac{h_n}{\theta(\theta + t_n h_n)} \to -\frac{h}{\theta^2},$$

where we have suppressed the argument in $g$ for clarity. Thus $\phi$ is Hadamard-differentiable, tangentially to $\mathbb{D}$, with $\phi'_\theta(h) = -h/\theta^2$.

Sometimes Hadamard differentiability is also called *compact differentiability*. Another important property of this kind of derivative is that it satisfies a chain rule, in that compositions of Hadamard-differentiable functions are also Hadamard-differentiable. Details on this and several other aspects of functional differentiation will be postponed until Part II. We have the following important result (the proof of which will be given in Part II, Page 235):

THEOREM 2.8 *For normed spaces $\mathbb{D}$ and $\mathbb{E}$, let $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ be Hadamard-differentiable at $\theta$ tangentially to $\mathbb{D}_0 \subset \mathbb{D}$. Assume that $r_n(X_n - \theta) \rightsquigarrow X$ for some sequence of constants $r_n \to \infty$, where $X_n$ takes its values in $\mathbb{D}_\phi$, and $X$ is a tight process taking its values in $\mathbb{D}_0$. Then $r_n(\phi(X_n) - \phi(\theta)) \rightsquigarrow \phi'_\theta(X)$.*

Consider again the quantile map $\phi_\xi$, and let the distribution function $F$ be absolutely continuous over $N = [u, v] = [F^{-1}(a) - \epsilon, F^{-1}(b) + \epsilon]$, for some $\epsilon > 0$, with continuous density $f$ such that $0 < \inf_{t\in N} f(t) \leq \sup_{t\in N} f(t) < \infty$. Also let $\mathbb{D}_1 \subset D[u, v]$ be the space of all distribution functions restricted to $[u, v]$. We will now argue that $\phi_\xi$ is Hadamard-differentiable at $F$ tangentially to $C[u, v]$, where for any interval $A \subset \mathbb{R}$, $CA$ is the space of continuous real functions on $A$. Let $t_n \to 0$ and $\{h_n\} \in D[u, v]$ converge uniformly to $h \in C[u, v]$ such that $F + t_n h_n \in \mathbb{D}_1$ for all $n \geq 1$, and denote $\xi_p = F^{-1}(p)$, $\xi_{pn} = (F + t_n h_n)^{-1}(p)$, $\xi_{pn}^N = (\xi_{pn} \vee u) \wedge v$, and $\epsilon_{pn} = t_n^2 \wedge (\xi_{pn}^N - u)$. The reason for the modification $\xi_{pn}^N$ is to ensure that the quantile estimate is contained in $[u, v]$ and hence also $\epsilon_{pn} \geq 0$. Thus there exists an $n_0 < \infty$, such that for all $n \geq n_0$, $(F + t_n h_n)(u) < a$, $(F + t_n h_n)(v) > b$, $\epsilon_{pn} > 0$ and $\xi_{pn}^N = \xi_{pn}$ for all $p \in [a, b]$, and therefore

$$(F + t_n h_n)(\xi_{pn}^N - \epsilon_{pn}) \leq F(\xi_p) \leq (F + t_n h_n)(\xi_{pn}^N) \tag{2.9}$$

for all $p \in [a, b]$, since $(F + t_n h_n)^{-1}(p)$ is the smallest $x$ satisfying $(F + t_n h_n)(x) \geq p$ and $F(\xi_p) = p$.

Since $F(\xi_{pn}^N - \epsilon_{pn}) = F(\xi_{pn}^N) + O(\epsilon_{pn})$, $h_n(\xi_{pn}^N) - h(\xi_{pn}^N) = o(1)$, and $h_n(\xi_{pn}^N - \epsilon_{pn}) - h(\xi_{pn}^N - \epsilon_{pn}) = o(1)$, where $O$ and $o$ are uniform over $p \in [a, b]$ (here and for the remainder of our argument), we have that (2.9) implies

$$\begin{aligned} F(\xi_{pn}^N) + t_n h(\xi_{pn}^N - \epsilon_{pn}) + o(t_n) &\leq F(\xi_p) \\ &\leq F(\xi_{pn}^N) + t_n h(\xi_{pn}^N) + o(t_n). \end{aligned} \tag{2.10}$$

But this implies that $F(\xi_{pn}^N) + O(t_n) \leq F(\xi_p) \leq F(\xi_{pn}^N) + O(t_n)$, which implies that $|\xi_{pn} - \xi_p| = O(t_n)$. This, together with (2.10) and the fact that $h$ is continuous, implies that $F(\xi_{pn}) - F(\xi_p) = -t_n h(\xi_p) + o(t_n)$. This now yields

$$\frac{\xi_{pn} - \xi_p}{t_n} = -\frac{h(\xi_p)}{f(\xi_p)} + o(1),$$

and the desired Hadamard-differentiability of $\phi_\xi$ follows, with derivative $\phi'_F(h) = \{-h(F^{-1}(p))/f(F^{-1}(p)), p \in [a, b]\}$.

The functional delta method also applies to the bootstrap. Consider the sequence of random elements $\mathbb{X}_n(X_n)$ in a normed space $\mathbb{D}$, and assume that $r_n(\mathbb{X}_n - \mu) \rightsquigarrow \mathbb{X}$, where $\mathbb{X}$ is tight in $\mathbb{D}$, for some sequence of constants $0 < r_n \rightsquigarrow \infty$. Here, $\mathbb{X}_n$ is a generic empirical process based on the data sequence $\{X_n, n \geq 1\}$, and is not restricted to i.i.d. data. Now assume we have a bootstrap of $\mathbb{X}_n$, $\hat{\mathbb{X}}_n(X_n, W_n)$, where $W = \{W_n\}$ is a sequence of random bootstrap weights which are independent of $X_n$. Also assume $\hat{\mathbb{X}}_n \overset{\mathrm{P}}{\underset{W}{\rightsquigarrow}} \mathbb{X}$. We have the following bootstrap result:

THEOREM 2.9 *For normed spaces $\mathbb{D}$ and $\mathbb{E}$, let $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ be Hadamard-differentiable at $\mu$ tangentially to $\mathbb{D}_0 \subset \mathbb{D}$, with derivative $\phi'_\mu$.*

*Let $\mathbb{X}_n$ and $\hat{\mathbb{X}}_n$ have values in $\mathbb{D}_\phi$, with $r_n(\mathbb{X}_n - \mu) \rightsquigarrow \mathbb{X}$, where $\mathbb{X}$ is tight and takes its values in $\mathbb{D}_0$, the maps $W_n \mapsto \hat{\mathbb{X}}_n$ are appropriately measurable, and where $r_n c(\hat{\mathbb{X}}_n - \mathbb{X}_n) \overset{\mathrm{P}}{\underset{W}{\rightsquigarrow}} \mathbb{X}$, for some $0 < c < \infty$. Then $r_n c(\phi(\hat{\mathbb{X}}_n) - \phi(\mathbb{X}_n)) \overset{\mathrm{P}}{\underset{W}{\rightsquigarrow}} \phi'_\mu(\mathbb{X})$.*

We will postpone until Part II a more precise discussion of what "appropriately measurable" means in this context (see Page 236).

When $\mathbb{X}_n$ in the previous theorem is the empirical process $\mathbb{P}_n$ indexed by a Donsker class $\mathcal{F}$ and $r_n = \sqrt{n}$, the results of Theorem 2.6 apply with $\mu = P$ for either the nonparametric or multiplier bootstrap weights. Moreover, the above measurability condition also holds (this will be verified in Chapter 12). Thus the bootstrap is automatically valid for Hadamard-differentiable functions applied to empirical processes indexed by Donsker classes. As a simple example, bootstraps of the quantile process $\{\hat{\xi}_n(p), p \in [a,b] \subset (0,1)\}$ are valid, provided the conditions given in the example following Theorem 2.8 for the density $f$ over the interval $N$ are satisfied. This can be used, for example, to create asymptotically valid confidence bands for $\{F^{-1}(p), p \in [a,b]\}$. There are also results for outer almost sure conditional convergence of the conditional laws of the bootstrapped process $r_n(\phi(\hat{\mathbb{X}}_n) - \phi(\mathbb{X}_n))$, but this requires stronger conditions on the differentiability of $\phi$, and we will not pursue this further in this book.

### *2.2.5 Z-Estimators*

A Z-estimator $\hat{\theta}_n$ is the approximate zero of a data-dependent function. To be more precise, let the parameter space be $\Theta$ and let $\Psi_n : \Theta \mapsto \mathbb{L}$ be a data-dependent function between two normed spaces, with norms $\|\cdot\|$ and $\|\cdot\|_{\mathbb{L}}$, respectively. If $\|\Psi_n(\hat{\theta}_n)\|_{\mathbb{L}} \overset{\mathrm{P}}{\to} 0$, then $\hat{\theta}_n$ is a Z-estimator. The main statistical issues for such estimators are consistency, asymptotic normality and validity of the bootstrap. Usually, $\Psi_n$ is an estimator of a fixed function $\Psi : \Theta \mapsto \mathbb{L}$ with $\Psi(\theta_0) = 0$ for some parameter of interest $\theta_0 \in \Theta$. We save the proof of the following theorem as an exercise:

THEOREM 2.10 *Let $\Psi(\theta_0) = 0$ for some $\theta_0 \in \Theta$, and assume $\|\Psi(\theta_n)\|_{\mathbb{L}} \to 0$ implies $\|\theta_n - \theta_0\| \to 0$ for any sequence $\{\theta_n\} \in \Theta$ (this is an "identifiability" condition). Then*

*(i) If $\|\Psi_n(\hat{\theta}_n)\|_{\mathbb{L}} \overset{\mathrm{P}}{\to} 0$ for some sequence of estimators $\hat{\theta}_n \in \Theta$ and $\sup_{\theta\in\Theta} \|\Psi_n(\theta) - \Psi(\theta)\|_{\mathbb{L}} \overset{\mathrm{P}}{\to} 0$, then $\|\hat{\theta}_n - \theta_0\| \overset{\mathrm{P}}{\to} 0$.*

*(ii) If $\|\Psi_n(\hat{\theta}_n)\|_{\mathbb{L}} \overset{\mathrm{as*}}{\to} 0$ for some sequence of estimators $\hat{\theta}_n \in \Theta$ and $\sup_{\theta\in\Theta} \|\Psi_n(\theta) - \Psi(\theta)\|_{\mathbb{L}} \overset{\mathrm{as*}}{\to} 0$, then $\|\hat{\theta}_n - \theta_0\| \overset{\mathrm{as*}}{\to} 0$.*

Consider, for example, estimating the survival function for right-censored failure time data. In this setting, we observe $X = (U, \delta)$, where $U = T \wedge C$, $\delta = 1\{T \leq C\}$, $T$ is a failure time of interest with distribution function $F_0$ and survival function $S_0 = 1 - F_0$ with $S_0(0) = 1$, and $C$ is a censoring time with distribution and survival functions $G$ and $L = 1 - G$, respectively, with $L(0) = 1$. For a sample of $n$ observations $\{X_i, i = 1, \ldots, n\}$, let $\{\tilde{T}_j, j = 1, \ldots, m_n\}$ be the unique observed failure times. The *Kaplan-Meier estimator* $\hat{S}_n$ of $S_0$ is then given by

$$\hat{S}_n(t) = \prod_{j:\tilde{T}_j \leq t} \left(1 - \frac{\sum_{i=1}^n \delta_i 1\{U_i = \tilde{T}_j\}}{\sum_{i=1}^n 1\{U_i \geq \tilde{T}_j\}}\right).$$

Consistency and other properties of this estimator can be demonstrated via standard continuous-time martingale arguments (Fleming and Harrington, 1991; Andersen, Borgun, Keiding and Gill, 1993); however, it is instructive to use empirical process arguments for Z-estimators.

Let $\tau < \infty$ satisfy $L(\tau-)S_0(\tau-) > 0$, and let $\Theta$ be the space of all survival functions $S$ with $S(0) = 1$ and restricted to $[0, \tau]$. We will use the uniform norm $\|\cdot\|_\infty$ on $\Theta$. After some algebra, the Kaplan-Meier estimator can be shown to be the solution of $\Psi_n(\hat{S}_n) = 0$, where $\Psi_n : \Theta \mapsto \Theta$ has the form $\Psi_n(S)(t) = \mathbb{P}_n \psi_{S,t}$, where

$$\psi_{S,t}(X) = 1\{U > t\} + (1 - \delta)1\{U \leq t\}1\{S(U) > 0\}\frac{S(t)}{S(U)} - S(t).$$

This is Efron's (1967) "self-consistency" expression for the Kaplan-Meier. For the fixed function $\Psi$, we use $\Psi(S)(t) = P\psi_{S,t}$. Somewhat surprisingly, the class of function $\mathcal{F} = \{\psi_{S,t} : S \in \Theta, t \in [0, \tau]\}$ is $P$-Donsker. To see this, first note that the class $\mathcal{M}$ of monotone functions $f : [0, \tau] \mapsto [0, 1]$ of the real random variable $U$ has bounded entropy (with bracketing) integral, which fact we establish later in Part II. Now the class of functions $\mathcal{M}_1 = \{\tilde{\psi}_{S,t} : S \in \Theta, t \in [0, \tau]\}$, where

$$\tilde{\psi}_{S,t}(U) = 1\{U > t\} + 1\{U \leq t\}1\{S(U) > 0\}\frac{S(t)}{S(U)},$$

is a subset of $\mathcal{M}$, since $\tilde{\psi}_{S,t}(U)$ is monotone in $U$ on $[0, \tau]$ and takes values only in $[0, 1]$ for all $S \in \Theta$ and $t \in [0, \tau]$. Note that $\{1\{U \leq t\} : t \in [0, \tau]\}$ is also Donsker (as argued previously), and so is $\{\delta\}$ (trivially) and $\{S(t) : S \in \Theta, t \in [0, \tau]\}$, since any class of fixed functions is always Donsker. Since all of these Donsker classes are bounded, we now have that $\mathcal{F}$ is Donsker since sums and products of bounded Donsker classes are also Donsker. Since Donsker classes are also Glivenko-Cantelli, we have that $\sup_{S \in \Theta} \|\Psi_n(S) - \Psi(S)\|_\infty \stackrel{\text{as*}}{\to} 0$. If we can establish the identifiability condition for $\Psi$, the outer almost sure version of Theorem 2.10 gives us that $\|\hat{S}_n - S_0\|_\infty \stackrel{\text{as*}}{\to} 0$.

After taking expectations, the function $\Psi$ can be shown to have the form

$$\Psi(S)(t) = P\psi_{S,t} = S_0(t)L(t) + \int_0^t \frac{S_0(u)}{S(u)} dG(u)S(t) - S(t). \tag{2.11}$$

Thus, if we make the substitution $\epsilon_n(t) = S_0(t)/S_n(t) - 1$, $\Psi(S_n)(t) \to 0$ uniformly over $t \in [0, \tau]$ implies that $u_n(t) = \epsilon_n(t)L(t) + \int_0^t \epsilon_n(u)dG(u) \to 0$ uniformly over the same interval. By solving this integral equation, we obtain $\epsilon_n(t) = u_n(t)/L(t-) - \int_0^{t-} [L(s)L(s-)]^{-1} u_n(s)dG(s)$, which implies $\epsilon_n(t) \to 0$ uniformly, since $L(t-) \geq L(\tau-) > 0$. Thus $\|S_n - S_0\|_\infty \to 0$, implying the desired identifiability.

We now consider weak convergence of Z-estimators. Let $\Psi_n$, $\Psi$, $\Theta$ and $\mathbb{L}$ be as at the beginning of this section. We have the following master theorem for Z-estimators, the proof of which will be given in Part II (Page 254):

THEOREM 2.11 *Assume that $\Psi(\theta_0) = 0$ for some $\theta_0$ in the interior of $\Theta$, $\sqrt{n}\Psi_n(\hat{\theta}_n) \xrightarrow{\mathrm{P}} 0$, and $\|\hat{\theta}_n - \theta_0\| \xrightarrow{\mathrm{P}} 0$ for the random sequence $\{\hat{\theta}_n\} \in \Theta$. Assume also that $\sqrt{n}(\Psi_n - \Psi)(\theta_0) \rightsquigarrow Z$, for some tight random $Z$, and that*

$$\frac{\left\| \sqrt{n}(\Psi_n(\hat{\theta}_n) - \Psi(\hat{\theta}_n)) - \sqrt{n}(\Psi_n(\theta_0) - \Psi(\theta_0)) \right\|_{\mathbb{L}}}{1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|} \xrightarrow{\mathrm{P}} 0. \tag{2.12}$$

*If $\theta \mapsto \Psi(\theta)$ is Fréchet-differentiable at $\theta_0$ (defined below) with continuously-invertible (also defined below) derivative $\dot{\Psi}_{\theta_0}$, then*

$$\|\sqrt{n}\dot{\Psi}_{\theta_0}(\hat{\theta}_n - \theta_0) + \sqrt{n}(\Psi_n - \Psi)(\theta_0)\|_{\mathbb{L}} \xrightarrow{\mathrm{P}} 0 \tag{2.13}$$

*and thus $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow -\dot{\Psi}_{\theta_0}^{-1}(Z)$.*

*Fréchet-differentiability* of a map $\phi : \Theta \subset \mathbb{D} \mapsto \mathbb{L}$ at $\theta \in \Theta$ is stronger than Hadamard-differentiability, in that it means there exists a continuous, linear map $\phi'_\theta : \mathbb{D} \mapsto \mathbb{L}$ with

$$\frac{\|\phi(\theta + h_n) - \phi(\theta) - \phi'_\theta(h_n)\|_{\mathbb{L}}}{\|h_n\|} \to 0 \tag{2.14}$$

for all sequences $\{h_n\} \subset \mathbb{D}$ with $\|h_n\| \to 0$ and $\theta + h_n \in \Theta$ for all $n \geq 1$. *Continuous invertibility* of an *operator* $A : \Theta \mapsto \mathbb{L}$ essentially means $A$ is invertible with the property that for a constant $c > 0$ and all $\theta_1, \theta_2 \in \Theta$,

$$\|A(\theta_1) - A(\theta_2)\|_{\mathbb{L}} \geq c\|\theta_1 - \theta_2\|. \tag{2.15}$$

An operator is a map between spaces of function, such as the maps $\Psi$ and $\Psi_n$. We will postpone further discussion of operators and continuous invertibility until Part II.

Returning to our Kaplan-Meier example, with $\Psi_n(S)(t) = \mathbb{P}_n\psi_{S,t}$ and $\Psi(S)(t) = P\psi_{S,t}$ as before, note that since $\mathcal{F} = \{\psi_{S,t}, S \in \Theta, t \in [0, \tau]\}$ is

Donsker, we easily have that $\sqrt{n}(\Psi_n - \Psi)(\theta_0) \rightsquigarrow Z$, for $\theta_0 = S_0$ and some tight random $Z$. We also have that for any $\{S_n\} \in \Theta$ converging uniformly to $S_0$,

$$\begin{aligned} \sup_{t\in[0,\tau]} P(\psi_{S_n,t} - \psi_{S_0,t})^2 &\leq 2 \sup_{t\in[0,\tau]} \int_0^t \left[\frac{S_n(t)}{S_n(u)} - \frac{S_0(t)}{S_0(u)}\right]^2 S_0(u)dG(u) \\ &\quad +2 \sup_{t\in[0,\tau]} (S_n(t) - S_0(t))^2 \\ &\to 0. \end{aligned}$$

This can be shown to imply (2.12). After some analysis, $\Psi$ can be shown to be Fréchet-differentiable at $S_0$, with derivative

$$\dot{\Psi}_{\theta_0}(h)(t) = -\int_0^t \frac{S_0(t)h(u)}{S_0(u)} dG(u) - L(t)h(t), \tag{2.16}$$

for all $h \in D[0,\tau]$, having continuous inverse

$$\begin{aligned} \dot{\Psi}^{-1}_{\theta_0}(a)(t) &= -S_0(t) \\ &\quad \times \left\{ a(0) + \int_0^t \frac{1}{L(u-)S_0(u-)} \left[ da(u) + \frac{a(u)dF_0(u)}{S_0(u)} \right] \right\}, \end{aligned} \tag{2.17}$$

for all $a \in D[0,\tau]$. Thus all of the conditions of Theorem 2.11 are satisfied, and we obtain the desired weak convergence of $\sqrt{n}(\hat{S}_n - S_0)$ to a tight, mean zero Gaussian process. The covariance of this process is

$$V(s,t) = S_0(s)S_0(t) \int_0^{s\wedge t} \frac{dF_0(u)}{L(u-)S_0(u)S_0(u-)},$$

which can be derived after lengthy but straightforward calculations (which we omit).

Returning to general Z-estimators, there are a number of methods for showing that the conditional law of a bootstrapped Z-estimator, given the observed data, converges to the limiting law of the original Z-estimator. One important approach that is applicable to non-i.i.d. data involves establishing Hadamard-differentiability of the map $\phi$, which extracts a zero from the function $\Psi$. We will explore this approach in Part II. We close this section with a simple bootstrap result for the setting where $\Psi_n(\theta)(h) = \mathbb{P}_n\psi_{\theta,h}$ and $\Psi(\theta)(h) = P\psi_{\theta,h}$, for random and fixed real maps indexed by $\theta \in \Theta$ and $h \in \mathcal{H}$. Assume that $\Psi(\theta_0)(h) = 0$ for some $\theta_0 \in \Theta$ and all $h \in \mathcal{H}$, that $\sup_{h\in\mathcal{H}} |\Psi(\theta_n)(h)| \to 0$ implies $\|\theta_n - \theta_0\| \to 0$ for any sequence $\{\theta_n\} \in \Theta$, and that $\Psi$ is Fréchet-differentiable with continuously invertible derivative $\dot{\Psi}'_{\theta_0}$. Also assume that $\mathcal{F} = \{\psi_{\theta,h} : \theta \in \Theta, h \in \mathcal{H}\}$ is $P$-G-C with $\sup_{\theta\in\Theta,h\in\mathcal{G}} P|\psi_{\theta,h}| < \infty$. Furthermore, assume that $\mathcal{G} = \{\psi_{\theta,h} : \theta \in \Theta, \|\theta - \theta_0\| \leq \delta, h \in \mathcal{H}\}$, where $\delta > 0$, is $P$-Donsker and that $\sup_{h\in\mathcal{H}} P(\psi_{\theta_n,h} - \psi_{\theta_0,h})^2 \to 0$ for any sequence $\{\theta_n\} \in \Theta$ with

$\|\theta_n - \theta_0\| \to 0$. Then, using arguments similar to those used in the Kaplan-Meier example and with the help of Theorems 2.10 and 2.11, we have that if $\hat{\theta}_n$ satisfies $\sup_{h\in\mathcal{H}} \left|\sqrt{n}\Psi_n(\hat{\theta}_n)\right| \xrightarrow{\mathrm{P}} 0$, then $\|\hat{\theta}_n - \theta_0\| \xrightarrow{\mathrm{P}} 0$ and $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow -\dot{\Psi}_{\theta_0}^{-1}(Z)$, where $Z$ is the tight limiting distribution of $\sqrt{n}(\Psi_n(\theta_0) - \Psi(\theta_0))$.

Let $\Psi_n^\circ(\theta)(h) = \mathbb{P}_n^\circ \psi_{\theta,h}$, where $\mathbb{P}_n^\circ$ is either the nonparametric bootstrap $\hat{\mathbb{P}}_n$ or the multiplier bootstrap $\tilde{\mathbb{P}}_n$ defined in Section 2.2.3, and define the bootstrap estimator $\hat{\theta}_n^\circ \in \Theta$ to be a minimizer of $\sup_{h\in\mathcal{H}} |\Psi_n^\circ(\theta)(h)|$ over $\theta \in \Theta$. We will prove in Part II that these conditions are more than enough to ensure that $\sqrt{n}(\hat{\theta}_n^\circ - \hat{\theta}_n) \overset{\mathrm{P}}{\underset{W}{\rightsquigarrow}} -\dot{\Psi}_{\theta_0}^{-1}(Z)$, where $W$ refers to either the nonparametric or multiplier bootstrap weights. Thus the bootstrap is valid. These conditions for the bootstrap are satisfied in the Kaplan-Meier example, for either the nonparametric or multiplier weights, thus enabling the construction of confidence bands for $S_0(t)$ over $t \in [0, \tau]$.

### *2.2.6 M-Estimators*

An M-estimator $\hat{\theta}_n$ is the approximate maximum of a data-dependent function. To be more precise, let the parameter set be a metric space $(\Theta, d)$ and let $M_n : \Theta \mapsto \mathbb{R}$ be a data-dependent real function. If $M_n(\hat{\theta}_n) = \sup_{\theta\in\Theta} M_n(\theta) - o_P(1)$, then $\hat{\theta}_n$ is an M-estimator. Maximum likelihood and least-squares (after changing the sign of the objective function) estimators are some of the most important examples, but there are many other examples as well. As with Z-estimators, the main statistical issues for M-estimators are consistency, weak convergence and validity of the bootstrap. Unlike Z-estimators, the rate of convergence for M-estimators is not necessarily $\sqrt{n}$, even for i.i.d. data, and finding the right rate can be quite challenging.

For establishing consistency, $M_n$ is often an estimator of a fixed function $M : \Theta \mapsto \mathbb{R}$. We now present the following consistency theorem (the proof of which is deferred to Part II, Page 267):

THEOREM 2.12 *Assume for some $\theta_0 \in \Theta$ that $\liminf_{n\to\infty} M(\theta_n) \geq M(\theta_0)$ implies $d(\theta_n, \theta_0) \to 0$ for any sequence $\{\theta_n\} \in \Theta$ (this is another identifiability condition). Then, for a sequence of estimators $\hat{\theta}_n \in \Theta$,*

*(i) If $M_n(\hat{\theta}_n) = \sup_{\theta\in\Theta} M_n(\theta) - o_P(1)$ and $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\mathrm{P}} 0$, then $d(\hat{\theta}_n, \theta_0) \xrightarrow{\mathrm{P}} 0$.*

*(ii) If $M_n(\hat{\theta}_n) = \sup_{\theta\in\Theta} M_n(\theta) - o_{as*}(1)$ and $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\mathrm{as*}} 0$, then $d(\hat{\theta}_n, \theta_0) \xrightarrow{\mathrm{as*}} 0$.*

Suppose, for now, we know that the rate of convergence for the M-estimator $\hat{\theta}_n$ is $r_n$, or, in other words, we know that $Z_n = r_n(\hat{\theta}_n - \theta_0) =$

$O_P(1)$. $Z_n$ can now be re-expressed as the approximate maximum of the criterion function $h \mapsto H_n(h) = M_n(\theta_0 + h/r_n)$ for $h$ ranging over some metric space $\mathbb{H}$. If the argmax of $H_n$ over bounded subsets of $\mathbb{H}$ can now be shown to converge weakly to the argmax of a tight limiting process $H$ over the same bounded subsets, then $Z_n$ converges weakly to $\text{argmax}_{h\in\mathbb{H}} H(h)$.

We will postpone the technical challenges associated with determining these rates of convergence until Part II, and restrict ourselves to an interesting special case involving Euclidean parameters, where the rate is known to be $\sqrt{n}$. The proof of the following theorem is also deferred to Part II (Page 270):

THEOREM 2.13 *Let $X_1, \ldots, X_n$ be i.i.d. with sample space $\mathcal{X}$ and law $P$, and let $m_\theta : \mathcal{X} \mapsto \mathbb{R}$ be measurable functions indexed by $\theta$ ranging over an open subset of Euclidean space $\Theta \subset \mathbb{R}^p$. Let $\theta_0$ be a bounded point of maximum of $Pm_\theta$ in the interior of $\Theta$, and assume for some neighborhood $\Theta_0 \subset \Theta$ including $\theta_0$, that there exists measurable functions $\dot{m} : \mathcal{X} \mapsto \mathbb{R}$ and $\dot{m}_{\theta_0} : \mathcal{X} \mapsto \mathbb{R}^p$ satisfying*

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq \dot{m}(x)\|\theta_1 - \theta_2\|, \tag{2.18}$$

$$P[m_\theta - m_{\theta_0} - (\theta - \theta_0)'\dot{m}_{\theta_0}]^2 = o(\|\theta - \theta_0\|^2), \tag{2.19}$$

*$P\dot{m}^2 < \infty$, and $P\|\dot{m}_{\theta_0}\|^2 < \infty$, for all $\theta_1, \theta_2, \theta \in \Theta_0$. Assume also that $M(\theta) = Pm_\theta$ admits a second order Taylor expansion with nonsingular second derivative matrix $V$. Denote $M_n(\theta) = \mathbb{P}_n m_\theta$, and assume the approximate maximizer $\hat{\theta}_n$ satisfies $M_n(\hat{\theta}_n) = \sup_{\theta\in\Theta} M_n(\theta) - o_P(n^{-1})$ and $\|\hat{\theta}_n - \theta_0\| \xrightarrow{\text{P}} 0$. Then $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow -V^{-1}Z$, where $Z$ is the limiting Gaussian distribution of $\mathbb{G}_n\dot{m}_{\theta_0}$.*

Consider, for example, least absolute deviation regression. In this setting, we have i.i.d. random vectors $U_1, \ldots, U_n$ in $\mathbb{R}^p$ and random errors $e_1, \ldots, e_n$, but we observe only the data $X_i = (Y_i, U_i)$, where $Y_i = \theta_0' U_i + e_i$, $i = 1, \ldots, n$. The least-absolute-deviation estimator $\hat{\theta}_n$ minimizes the function $\theta \mapsto \mathbb{P}_n\tilde{m}_\theta$, where $\tilde{m}_\theta(X) = |Y - \theta' U|$. Since a minimizer of a criterion function $M_n$ is also a maximizer of $-M_n$, M-estimation methods can be used in this context with only a change in sign. Although boundedness of the parameter space $\Theta$ is not necessary for this regression setting, we restrict—for ease of discourse—$\Theta$ to be a bounded, open subset of $\mathbb{R}^p$ containing $\theta_0$. We also assume that the distribution of the errors $e_i$ has median zero and positive density at zero, which we denote $f(0)$, and that $P[UU']$ is finite and positive definite.

Note that since we are not assuming $\text{E}|e_i| < \infty$, it is possible that $P\tilde{m}_\theta = \infty$ for all $\theta \in \Theta$. Since minimizing $\mathbb{P}_n\tilde{m}_\theta$ is the same as minimizing $\mathbb{P}_n m_\theta$, where $m_\theta = \tilde{m}_\theta - \tilde{m}_{\theta_0}$, we will use $M_n(\theta) = \mathbb{P}_n m_\theta$ as our criterion function hereafter (without modifying the estimator $\hat{\theta}_n$). By the definition of $Y$, $m_\theta(X) = |e - (\theta - \theta_0)'U| - |e|$, and we now have that

$Pm_\theta \leq \|\theta - \theta_0\| \left(\mathrm{E}\|U\|^2\right)^{1/2} < \infty$ for all $\theta \in \Theta$. Since

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \quad \leq \quad \|\theta_1 - \theta_2\| \times \|u\|, \tag{2.20}$$

it is not hard to show that the class of function $\{m_\theta : \theta \in \Theta\}$ is $P$-Glivenko-Cantelli. It can also be shown that $Pm_\theta \geq 0$ with equality only when $\theta = \theta_0$. Hence Theorem 2.12, Part (ii), yields that $\hat{\theta}_n \overset{\text{as*}}{\to} \theta_0$.

Now we consider $M(\theta) = Pm_\theta$. By conditioning on $U$, one can show after some analysis that $M(\theta)$ is two times continuously differentiable, with second derivative $V = 2f(0)P[UU']$ at $\theta_0$. Note that (2.20) satisfies Condition (2.18); and with $\dot{m}_\theta(X) = -U\mathrm{sign}(e)$, we also have that

$$|m_\theta(X) - m_{\theta_0}(X) - (\theta - \theta_0)'\dot{m}_\theta(X)| \leq 1\{|e| \leq |(\theta - \theta_0)'U|\}\left[(\theta - \theta_0)'U\right]^2$$

satisfies Condition (2.19). Thus all the conditions of Theorem 2.13 are satisfied. Hence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically mean zero normal, with variance $V^{-1}P\left[\dot{m}_{\theta_0}\dot{m}'_{\theta_0}\right]V^{-1} = \left(P[UU']\right)^{-1}/\left(4f^2(0)\right)$. This variance is not difficult to estimate from the data, but we postpone presenting the details.

Another technique for obtaining weak convergence of M-estimators that are $\sqrt{n}$ consistent, is to first establish consistency and then take an appropriate derivative of the criterion function $M_n(\theta)$, $\Psi_n(\theta)(h)$, for $h$ ranging over some index set $H$, and apply Z-estimator techniques to $\Psi_n$. This works because the derivative of a smooth criterion function at an approximate maximizer is approximately zero. This approach facilitates establishing the validity of the bootstrap since such validity is often easier to obtain for Z-estimators than for M-estimators. This approach is also applicable to certain nonparametric maximum likelihood estimators which we will consider in Part III.

## 2.3 Other Topics

In addition to the empirical process topics outlined in the previous sections, we will cover a few other related topics in Part II, including results for sums of independent but not identically distributed stochastic processes and, briefly, for dependent but stationary processes. However, there are a number of interesting empirical process topics we will not pursue in later chapters, including general results for convergence of nets. In the remainder of this section, we briefly outline a few additional topics not covered later which involve sequences of empirical processes based on i.i.d. data. For some of these topics, we will primarily restrict ourselves to the empirical process $G_n = \sqrt{n}(\mathbb{F}_n - F)$, although many of these results have extensions which apply to more general empirical processes.

The law of the iterated logarithm for $G_n$ states that

$$\limsup_{n\to\infty} \frac{\|G_n\|_\infty}{\sqrt{2\log\log n}} \leq \frac{1}{2}, \quad \text{a.s.}, \tag{2.21}$$

with equality if 1/2 is in the range of $F$, where $\|\cdot\|_\infty$ is the uniform norm. This can be generalized to empirical processes on $P$-Donsker classes $\mathcal{F}$ which have a measurable envelope with bounded second moment (Dudley and Philipp, 1983):

$$\limsup_{n\to\infty} \frac{\left[\sup_{f\in\mathcal{F}} |\mathbb{G}_n(f)|\right]^*}{\sqrt{(2\log\log n)\sup_{f\in\mathcal{F}} |P(f-Pf)^2|}} \leq 1, \quad \text{a.s.}$$

Result (2.21) can be further strengthened to Strassen's (1964) theorem, which states that on a set with probability 1, the set of all limiting paths of $\sqrt{1/(2\log\log n)}G_n$ is exactly the set of all functions of the form $h(F)$, where $h(0) = h(1) = 0$ and $h$ is absolutely continuous with derivative $h'$ satisfying $\int_0^1 [h'(s)]^2\, ds \leq 1$. While the previous results give upper bounds on $\|G_n\|_\infty$, it is also known that

$$\liminf_{n\to\infty} \sqrt{2\log\log n}\|G_n\|_\infty = \frac{\pi}{2}, \quad \text{a.s.},$$

implying that the smallest uniform distance between $\mathbb{F}_n$ and $F$ is at least $O(1/\sqrt{n\log\log n})$.

A topic of interest regarding Donsker theorems is the closeness of the empirical process sample paths to the limiting Brownian bridge sample paths. The strongest result on this question for the empirical process $G_n$ is the KMT construction, named after Komlós, Major and Tusnády (1975, 1976). The KMT construction states that there exists fixed positive constants $a$, $b$, and $c$, and a sequence of standard Brownian bridges $\{\mathbb{B}_n\}$, such that

$$\mathrm{P}\left(\|G_n - \mathbb{B}_n(F)\|_\infty > \frac{a\log n + x}{\sqrt{n}}\right) \leq be^{-cx},$$

for all $x > 0$ and $n \geq 1$. This powerful result can be shown to imply both

$$\begin{aligned} \limsup_{n\to\infty} \frac{\sqrt{n}}{\log n}\|G_n - \mathbb{B}_n(F)\|_\infty &< \infty, \quad \text{a.s., and} \\ \limsup_{n\to\infty} \mathrm{E}\left[\frac{\sqrt{n}}{\log n}\|G_n - \mathbb{B}_n(F)\|_\infty\right]^m &< \infty, \end{aligned}$$

for all $0 < m < \infty$. These results are called *strong approximations* and have applications in statistics, such as in the construction of confidence bands for kernel density estimators (see, for example, Bickel and Rosenblatt, 1973). Another interesting application—to "large p, small n" asymptotics for

microarrays—will be developed in some detail in Section 15.5 of Part II, although we will not address the theoretical derivation of the KMT construction.

An important class of generalizations of the empirical process for i.i.d. data are the U-processes. The $m$th order empirical U-process measure $\mathbb{U}_{n,m}$ is defined, for a measurable function $f : \mathcal{X}^m \mapsto \mathbb{R}$ and a sample of obervations $X_1, \ldots, X_n$ on $\mathcal{X}$, as

$$\binom{n}{m}^{-1} \sum_{(i_1,\ldots,i_m)\in I_{n,m}} f(X_{i_1}, \ldots, X_{i_m}),$$

where $I_{n,m}$ is the set of all $m$-tuples of integers $(i_1, \ldots, i_m)$ satisfying $1 \leq i_1 < \cdots < i_m \leq n$. For $m = 1$, this measure reduces to the usual empirical measure, i.e., $\mathbb{U}_{n,1} = \mathbb{P}_n$. The empirical U-process, for a class of $m$-variate real functions $\mathcal{F}$ (of the form $f : \mathcal{X}^m \mapsto \mathbb{R}$ as above), is

$$\left\{\sqrt{n}(\mathbb{U}_{n,m} - P)f : f \in \mathcal{F}\right\}.$$

These processes are useful for solving a variety of complex statistical problems arising in a number of areas, including nonparametric monotone regression (see Ghosal, Sen and van der Vaart, 2000), testing for normality (see Arcones and Wang, 2006), and a number of other areas. Fundamental work on Glivenko-Cantelli and Donsker type results for U-processes can be found in Nolan and Pollard (1987, 1988) and Arcones and Giné (1993), and some useful technical tools can be found in Giné (1997). Recent developments include Monte Carlo methods of inference for U-processes (see, for example, Zhang, 2001) and a central limit theorem for two-sample U-processes (Neumeyer, 2004).

## 2.4 Exercises

**2.4.1.** Let $X, Y$ be a pair of real random numbers with joint distribution $P$. Compute upper bounds for $N_{[]}(\epsilon, \mathcal{F}, L_r(P))$, for $r = 1, 2$, where $\mathcal{F} = \{1\{X \leq s, Y \leq t\} : s, t \in \mathbb{R}\}$.

**2.4.2.** Prove Theorem 2.10.

**2.4.3.** Consider the Z-estimation framework for the Kaplan-Meier estimator discusses in Section 2.2.5. Let $\Psi(S)(t)$ be as defined in (2.11). Show that $\Psi$ is Fréchet-differentiable at $S_0$, with derivative $\dot{\Psi}_{\theta_0}(h)(t)$ given by (2.16), for all $h \in D[0, \tau]$.

**2.4.4.** Continuing with the set-up of the previous problem, show that $\dot{\Psi}_{\theta_0}$ is continuously invertible, with inverse $\dot{\Psi}^{-1}_{\theta_0}$ given in (2.17). The following approach may be easiest: First show that for any $a \in D[0, \tau]$,

$h(t) = \dot{\Psi}_{\theta_0}^{-1}(a)(t)$ satisfies $\dot{\Psi}_{\theta_0}(h)(t) = a(t)$. The following identity may be helpful:

$$d\left[\frac{a(t)}{S_0(t)}\right] = \frac{da(t)}{S_0(t-)} + \frac{a(t)dF_0(t)}{S_0(t-)S_0(t)}.$$

Now show that there exists an $M < \infty$ such that $\left\|\dot{\Psi}_{\theta_0}^{-1}(a)\right\| \leq M\|a\|$, where $\|\cdot\|$ is the uniform norm. This then implies that there exists a $c > 0$ such that $\|\dot{\Psi}_{\theta_0}(h)\| \geq c\|h\|$.

## 2.5 Notes

Theorem 2.1 is a composite of Theorems 1.5.4 and 1.5.7 of van der Vaart and Wellner (1996) (hereafter abbreviated VW). Theorems 2.2, 2.3, 2.4 and 2.5 correspond to Theorems 19.4, 19.5, 19.13 and 19.14, respectively, of van der Vaart (1998). The if and only if implications of (2.8) are described in VW, Page 73. The implications (i)⇔(ii) in Theorems 2.6 and 2.7 are given in Theorems 3.6.1 and 3.6.2, respectively, of VW. Theorems 2.8 and 2.11 correspond to Theorems 3.9.4 and 3.3.1, of VW, while Theorem 2.13 comes from Example 3.2.22 of VW.

# 3
# Overview of Semiparametric Inference

This chapter presents an overview of the main ideas and techniques of semiparametric inference, with particular emphasis on semiparametric efficiency. The major distinction between this kind of efficiency and the standard notion of efficiency for parametric maximum likelihood estimators—as expressed in the Cramér-Rao lower bound—is the presence of an infinite-dimensional nuisance parameter in semiparametric models. Proofs and other technical details will generally be postponed until Part III.

In the first section, we define and sketch the main features of semiparametric models and semiparametric efficiency. The second section discusses efficient score functions and estimating equations and their connection to efficient estimation. The third section discusses nonparametric maximum likelihood estimation, the main tool for constructing efficient estimators. The fourth and final section briefly discusses several additional related topics, including variance estimation and confidence band construction for efficient estimators.

## 3.1 Semiparametric Models and Efficiency

A *statistical model* is a collection of probability measures $\{P \in \mathcal{P}\}$ on a sample space $\mathcal{X}$. Such models can be expressed in the form $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where $\Theta$ is some parameter space. *Semiparametric models* are statistical models where $\Theta$ has one or more infinite-dimensional component. For example, the parameter space for the linear regression model (1.1), where

$Y = \beta' Z + e$, consists of two components, a subset of $p$-dimensional Euclidean space (for the regression parameter $\beta$) and the infinite-dimensional space of all joint distribution functions of $(e, Z)$ with $\mathrm{E}[e|Z] = 0$ and $E[e^2|Z] \leq K$ almost surely, for some $K < \infty$.

The goal of *semiparametric inference* is to construct optimal estimators and test statistics for evaluating semiparametric model parameters. The parameter component of interest can be succinctly expressed as a function of the form $\psi : \mathcal{P} \mapsto \mathbb{D}$, where $\psi$ extracts the component of interest and takes values in $\mathbb{D}$. For now, we assume $\mathbb{D}$ is finite dimensional. As an illustration, if we are interested in the unconditional variance of the residual errors in the regression model (1.1), $\psi$ would be the map $\psi(P) = \int_{\mathbb{R}} e^2 dF(e)$, where $F(t) = P[e \leq t]$ is the unconditional residual distribution component of $P$. Throughout this book, the statistics of interest will be based on an i.i.d. sample, $X_1, \ldots, X_n$, of realizations from some $P \in \mathcal{P}$.

An estimator $T_n$ of the parameter $\psi(P)$, based on such a sample, is *efficient* if the limiting variance $V$ of $\sqrt{n}(T_n - \psi(P))$ is the smallest possible among all *regular* estimators of $\psi(P)$. The inverse of $V$ is the *information* for $T_n$. Regularity will be defined more explicitly later in this section, but suffice it to say for now that the limiting distribution of $\sqrt{n}(T_n - \psi(P))$ (as $n \to \infty$), for a regular estimator $T_n$, is continuous in $P$. Note that as we are changing $P$, the distribution of $T_n$ changes as well as the parameter $\psi(P)$. Not all estimators are regular, but most commonly used estimators in statistics are. Optimality of test statistics is closely related to efficient estimation, in that the most powerful test statistics for a hypothesis about a parameter are usually based on efficient estimators for that parameter.

The optimal efficiency for estimators of a parameter $\psi(P)$ depends in part on the complexity of the model $\mathcal{P}$. Estimation under the model $\mathcal{P}$ is more taxing than estimation under any parametric submodel $\mathcal{P}_0 = \{P_\theta : \theta \in \Theta_0\} \subset \mathcal{P}$, where $\Theta_0$ is finite dimensional. Thus the information for estimation under model $\mathcal{P}$ is worse than the information under any parametric submodel $\mathcal{P}_0$. If the information for the regular estimator $T_n$ is equal to the minimum of the information over all efficient estimators for all parametric submodels $\mathcal{P}_0$, then $T_n$ is *semiparametric efficient*. For semiparametric models, this minimizer is the best possible, since the only models with more information are parametric models. A parametric model that achieves this minimum, if such a model exists, is called a *least favorable* or *hardest* submodel. Note that efficient estimators for parametric models are trivially semiparametric efficient since such models are their own parametric submodels.

Fortunately, finding the minimum information over parametric submodels usually only requires consideration of one-dimensional parametric submodels $\{P_t : t \in N_\epsilon\}$ surrounding representative distributions $P \in \mathcal{P}$, where $N_\epsilon = [0, \epsilon)$ for some $\epsilon > 0$, $P_0 = P$, and $P_t \in \mathcal{P}$ for all $t \in N_\epsilon$. If $\mathcal{P}$ has a dominating measure $\mu$, then each $P \in \mathcal{P}$ can be expressed as a density $p$. In this case, we require the submodels around a representative density $p$

to be smooth enough so that the real function $g(x) = \partial \log p_t(x)/(\partial t)|_{t=0}$ exists with $\int_{\mathcal{X}} g^2(x)p(x)\mu(dx) < \infty$. This idea can be restated in sufficient generality to allow for models that may not be dominated. In this more general case, we require

$$\int \left[ \frac{(dP_t(x))^{1/2} - (dP(x))^{1/2}}{t} - \frac{1}{2} g(x)(dP(x))^{1/2} \right]^2 \to 0, \tag{3.1}$$

as $t \downarrow 0$. In this setting, we say that the submodel $\{P_t : t \in N_\epsilon\}$ is *differentiable in quadratic mean* at $t = 0$, with *score function* $g : \mathcal{X} \mapsto \mathbb{R}$.

In evaluating efficiency, it is necessary to consider many such one-dimensional submodels surrounding the representative $P$, each with a different score function. Such a collection of score functions is called a *tangent set* of the model $\mathcal{P}$ at $P$, and is denoted $\dot{\mathcal{P}}_P$. Because $Pg = 0$ and $Pg^2 < \infty$ for any $g \in \dot{\mathcal{P}}_P$, such tangent sets are subsets of $L_2^0(P)$, the space of all functions $h : \mathcal{X} \mapsto \mathbb{R}$ with $Ph = 0$ and $Ph^2 < \infty$. Note that, as we have done here, we will sometimes omit function arguments for simplicity, provided the context is clear. When the tangent set is closed under linear combinations, it is called a *tangent space*. Usually, one can take the closed linear span (the closure under linear combinations) of a tangent set to make a tangent space.

Consider $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where $\Theta$ is an open subset of $\mathbb{R}^k$. Assume that $\mathcal{P}$ is dominated by $\mu$, and that the classical score function $\dot{\ell}_\theta(x) = \partial \log p_\theta(x)/(\partial\theta)$ exists with $P_\theta[\dot{\ell}_\theta \dot{\ell}_\theta']$ bounded and $P_\theta \|\dot{\ell}_{\tilde{\theta}} - \dot{\ell}_\theta\|^2 \to 0$ as $\tilde{\theta} \to \theta$. Now for each $h \in \mathbb{R}^k$, let $\epsilon > 0$ be small enough so that $\{P_t : t \in N_\epsilon\} \subset \mathcal{P}$, where for each $t \in N_\epsilon$, $P_t = P_{\theta+th}$. One can show (see Exercise 3.5.1) that each of these one-dimensional submodels satisfy (3.1), with $g = h'\dot{\ell}_\theta$, resulting in the tangent space $\dot{\mathcal{P}}_{P_\theta} = \{h'\dot{\ell}_\theta : h \in \mathbb{R}^k\}$. Thus there is a simple connection between the classical score function and the more general idea of tangent sets and tangent spaces.

Continuing with the parametric setting, the estimator $\hat{\theta}$ is efficient for estimating $\theta$ if it is regular with information achieving the Cramér-Rao lower bound $P[\dot{\ell}_\theta \dot{\ell}_\theta']$. Thus the tangent set for the model contains information about the optimal efficiency. This is also true for semiparametric models in general, although the relationship between tangent sets and the optimal information is more complex.

Consider estimation of the parameter $\psi(P) \in \mathbb{R}^k$ for the semiparametric model $\mathcal{P}$. For any estimator $T_n$ of $\psi(P)$, if $\sqrt{n}(T_n - \psi(P)) = \sqrt{n}\mathbb{P}_n \breve{\psi}_P + o_P(1)$, where $o_P(1)$ denotes a quantity going to zero in probability, then $\breve{\psi}_P$ is an *influence function* for $\psi(P)$ and $T_n$ is *asymptotically linear*. For a given tangent set $\dot{\mathcal{P}}_P$, assume for each submodel $\{P_t : t \in N_\epsilon\}$ satisfying (3.1) with some $g \in \dot{\mathcal{P}}_P$ and some $\epsilon > 0$, that $d\psi(P_t)/(dt)|_{t=0} = \dot{\psi}_P(g)$ for some linear map $\dot{\psi}_P : L_2^0(P) \mapsto \mathbb{R}^k$. In this setting, we say that $\psi$ is differentiable at $P$ relative to $\dot{\mathcal{P}}_P$. When $\dot{\mathcal{P}}_P$ is a linear space, there exists a measurable function $\tilde{\psi}_P : \mathcal{X} \mapsto \mathbb{R}^k$ such that

$\dot{\psi}_P(g) = P\left[\tilde{\psi}_P(X)g(X)\right]$, for each $g \in \dot{\mathcal{P}}_P$. The function $\tilde{\psi}_P \in \dot{\mathcal{P}}_P \subset L_2^0(P)$ is unique and is called the *efficient influence function* for the parameter $\psi$ in the model $P$ (relative to the tangent space $\dot{\mathcal{P}}_P$). Here, and throughout the book, we abuse notation slightly by declaring that a random vector is in a given linear space if and only if each component of the vector is. Note that $\tilde{\psi}_P \in \dot{\mathcal{P}}_P$. Frequently, the efficient influence function can be found by taking a candidate influence function $\dot{\psi}_P \in L_2^0(P)$ and projecting it onto $\dot{\mathcal{P}}_P$ to obtain $\tilde{\psi}_P$. If $\sqrt{n}(T_n - \psi(P))$ is asymptotically equivalent to $\sqrt{n}\mathbb{P}_n\tilde{\psi}_P$, then $T_n$ can be shown to be semiparametric efficient (which we refer to hereafter simply as "efficient"). We will see in Theorem 18.8 in Section 18.2 (on Page 339), that full efficiency of an estimator $T_n$ can essentially be verified by checking whether the influence function for $T_n$ lies in $\dot{\mathcal{P}}_P$. This can be a relatively straightforward method of verification for many settings.

Consider again the parametric example, with $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^k$. Suppose the parameter of interest is $\psi(P_\theta) = f(\theta)$, where $f : \mathbb{R}^k \mapsto \mathbb{R}^d$ has derivative $\dot{f}_\theta$ at $\theta$, and that the Fisher information matrix $I_\theta = P[\dot{\ell}_\theta \dot{\ell}_\theta']$ is invertible. It is not hard to show that $\dot{\psi}_P(g) = \dot{f}_\theta I_\theta^{-1} P\left[\dot{\ell}_\theta(X)g(X)\right]$, and thus $\tilde{\psi}_P(x) = \dot{f}_\theta I_\theta^{-1}\dot{\ell}_\theta(x)$. Any estimator $T_n$ for which $\sqrt{n}(T_n - f(\theta))$ is asymptotically equivalent to $\sqrt{n}\mathbb{P}_n\left[\dot{f}_\theta I_\theta^{-1}\dot{\ell}_\theta\right]$, has asymptotic variance equal to the Cramér-Rao lower bound $\dot{f}_\theta I_\theta^{-1}\dot{f}_\theta'$.

Returning to the semiparametric setting, any one-dimensional submodel $\{P_t : t \in N_\epsilon\}$, satisfying (3.1) for the score function $g \in \dot{\mathcal{P}}_P$ and some $\epsilon > 0$, is a parametric model with parameter $t$. The Fisher information for $t$, evaluated at $t = 0$, is $Pg^2$. Thus the Cramér-Rao lower bound for estimating a univariate parameter $\psi(P)$ based on this model is $\left(P\left[\tilde{\psi}_P g\right]\right)^2 / Pg^2$, since $d\psi(P_t)/(dt)|_{t=0} = P\left[\tilde{\psi}_P g\right]$. Provided that $\tilde{\psi}_P \in \dot{\mathcal{P}}_P$ and all necessary derivatives exist, the maximum Cramér-Rao lower bound over all such submodels in $\dot{\mathcal{P}}_P$ is thus $P\tilde{\psi}_P^2$. For more general Euclidean parameters $\psi(P)$, this lower bound on the asymptotic variance is $P\left[\tilde{\psi}_P\tilde{\psi}_P'\right]$.

Hence, for $P\left[\tilde{\psi}_P\tilde{\psi}_P'\right]$ to be the upper bound for all parametric submodels, the tangent set must be sufficiently large. Obviously, the tangent set must also be restricted to score functions which reflect valid submodels. In addition, the larger the tangent set, the fewer the number of regular estimators. To see this, we will now provide a more precise definition of regularity. Let $P_{t,g}$ denote a submodel $\{P_t : t \in N_\epsilon\}$ satisfying (3.1) for the score $g$ and some $\epsilon > 0$. $T_n$ is regular for $\psi(P)$ if the limiting distribution of $\sqrt{n}(T_n - \psi(P_{1/\sqrt{n},g}))$, over the sequence of distributions $P_{1/\sqrt{n},g}$, exists and is constant over all $g \in \dot{\mathcal{P}}_P$. Thus the tangent set must be chosen to be large enough but not too large. Fortunately, most estimators in

common use for semiparametric inference are regular for large tangent sets, and thus there is usually quite a lot to be gained by making the effort to obtain an efficient estimator. Once the tangent set has been identified, the corresponding efficient estimator $T_n$ of $\psi(P)$ will always be asymptotically linear with influence function equal to the efficient influence function $\tilde{\psi}_P$.

Consider, for example, the unrestricted model $\mathcal{P}$ of all distributions on $\mathcal{X}$. Suppose we are interested in estimating $\psi(P) = Pf$ for some $f \in L_2(P)$, the space of all measurable functions $h$ with $Ph^2 < \infty$. For bounded $g \in L_2^0(P)$, the one-dimensional submodel $\{P_t : dP_t = (1 + tg)dP, t \in N_\epsilon\} \subset \mathcal{P}$ for $\epsilon$ small enough. Furthermore, (3.1) is satisfied with $\partial\psi(P_t)/(\partial t)|_{t=0} = P[fg]$. It is not hard to show, in fact, that one-dimensional submodels satisfying (3.1) with $\partial\psi(P_t)/(\partial t)|_{t=0} = P[fg]$ exist for all $g \in L_2^0(P)$. Thus $\dot{\mathcal{P}}_P = L_2^0(P)$ is a tangent set for the unrestricted model and $\tilde{\psi}_P(x) = f(x) - Pf$ is the corresponding efficient influence function. Since $\sqrt{n}(\mathbb{P}_n f - \psi(P)) = \sqrt{n}\mathbb{P}_n\tilde{\psi}_P$, $\mathbb{P}_n f$ is efficient for estimating $Pf$. Note that, in this unrestricted model, $\dot{\mathcal{P}}_P$ is the maximal possible tangent set, since all tangent sets must be subsets of $L_2^0(P)$. In general, the size of a tangent set reflects the amount of restrictions placed on a model, in that larger tangent sets reflect fewer restrictions.

Efficiency can also be established for infinite dimensional parameters $\psi(P)$ when $\sqrt{n}$ consistent regular estimators for $\psi(P)$ exist. There are a number of ways of expressing efficiency in this context, but we will only mention the convolution approach here. The convolution theorem states that for any regular estimator $T_n$ of $\psi(P)$, $\sqrt{n}(T_n - \psi(P))$ has a weak limiting distribution which is the convolution of a Gaussian process $Z$ and an independent process $M$, where $Z$ has the same limiting distribution as $\sqrt{n}\mathbb{P}_n\tilde{\psi}_P$. In other words, an inefficient estimator always has an asymptotically non-negligible independent noise process $M$ added to the efficient estimator distribution. A regular estimator $T_n$ for which $M$ is zero is efficient. Occasionally, we will use the term *uniformly efficient* when it is helpful to emphasize the fact that $\psi(P)$ is infinite dimensional. If $\psi(P)$ is indexed by $\{t \in T\}$, and if $T_n(t)$ is efficient for $\psi(P)(t)$ for each $t \in T$, then it can be shown that weak convergence of $\sqrt{n}(T_n - \psi(P))$ to a tight, mean zero Gaussian process implies uniform efficiency of $T_n$. This result will be formally stated in Theorem 18.9 on Page 341. Another important fact is that if $T_n$ is an efficient estimator for $\psi(P)$ and $\phi$ is a suitable Hadamard-differentiable function, then $\phi(T_n)$ is an efficient estimator for $\phi(\psi(P))$. We will make these results more explicit in Part III.

## 3.2 Score Functions and Estimating Equations

A parameter $\psi(P)$ of particular interest is the parametric component $\theta$ of a semiparametric model $\{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$, where $\Theta$ is an open

subset of $\mathbb{R}^k$ and $H$ is an arbitrary set that may be infinite dimensional. Tangent sets can be used to develop an efficient estimator for $\psi(P_{\theta,\eta}) = \theta$ through the formation of an *efficient score function*. In this setting, we consider submodels of the form $\{P_{\theta+ta,\eta_t}, t \in N_\epsilon\}$ that are differentiable in quadratic mean with score function $\partial \log dP_{\theta+ta,\eta_t}/(\partial t)|_{t=0} = a'\dot{\ell}_{\theta,\eta} + g$, where $a \in \mathbb{R}^k$, $\dot{\ell}_{\theta,\eta} : \mathcal{X} \mapsto \mathbb{R}^k$ is the ordinary score for $\theta$ when $\eta$ is fixed, and where $g : \mathcal{X} \mapsto \mathbb{R}$ is an element of a tangent set $\dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}}$ for the submodel $\mathcal{P}_\theta = \{P_{\theta,\eta} : \eta \in H\}$ (holding $\theta$ fixed). This tangent set is the *tangent set for* $\eta$ and should be rich enough to reflect all parametric submodels of $\mathcal{P}_\theta$. The tangent set for the full model is $\dot{\mathcal{P}}_{P_{\theta,\eta}} = \left\{a'\dot{\ell}_{\theta,\eta} + g : a \in \mathbb{R}^k, g \in \dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}}\right\}$.

While $\psi(P_{\theta+ta,\eta_t}) = \theta + ta$ is clearly differentiable with respect to $t$, we also need, as in the previous section, that there there exists a function $\tilde{\psi}_{\theta,\eta} : \mathcal{X} \mapsto \mathbb{R}^k$ such that

$$\left.\frac{\partial \psi(P_{\theta+ta,\eta_t})}{\partial t}\right|_{t=0} = a = P\left[\tilde{\psi}_{\theta,\eta}\left(\dot{\ell}'_{\theta,\eta}a + g\right)\right], \tag{3.2}$$

for all $a \in \mathbb{R}^k$ and all $g \in \dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}}$. After setting $a = 0$, we see that such a function must be uncorrelated with all of the elements of $\dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}}$.

Define $\Pi_{\theta,\eta}$ to be the orthogonal projection onto the closed linear span of $\dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}}$ in $L_2^0(P_{\theta,\eta})$. We will describe how to obtain such projections in detail in Part III, but suffice it to say that for any $h \in L_2^0(P_{\theta,\eta})$, $h = h - \Pi_{\theta,\eta}h + \Pi_{\theta,\eta}h$, where $\Pi_{\theta,\eta}h \in \dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}}$ but $P[(h - \Pi_{\theta,\eta}h)\, g] = 0$ for all $g \in \dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}}$. The *efficient score function* for $\theta$ is $\tilde{\ell}_{\theta,\eta} = \dot{\ell}_{\theta,\eta} - \Pi_{\theta,\eta}\dot{\ell}_{\theta,\eta}$, while the *efficient information matrix* for $\theta$ is $\tilde{I}_{\theta,\eta} = P\left[\tilde{\ell}_{\theta,\eta}\tilde{\ell}'_{\theta,\eta}\right]$.

Provided that $\tilde{I}_{\theta,\eta}$ is nonsingular, the function $\tilde{\psi}_{\theta,\eta} = \tilde{I}^{-1}_{\theta,\eta}\tilde{\ell}_{\theta,\eta}$ satisfies (3.2) for all $a \in \mathbb{R}^k$ and all $g \in \dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}}$. Thus the functional (parameter) $\psi(P_{\theta,\eta}) = \theta$ is differentiable at $P_{\theta,\eta}$ relative to the tangent set $\dot{\mathcal{P}}_{P_{\theta,\eta}}$, with efficient influence function $\tilde{\psi}_{\theta,\eta}$. Hence the search for an efficient estimator of $\theta$ is over if one can find an estimator $T_n$ satisfying $\sqrt{n}(T_n - \theta) = \sqrt{n}\mathbb{P}_n\tilde{\psi}_{\theta,\eta} + o_P(1)$. Note that $\tilde{I}_{\theta,\eta} = I_{\theta,\eta} - P\left[\Pi_{\theta,\eta}\dot{\ell}_{\theta,\eta}\left(\Pi_{\theta,\eta}\dot{\ell}_{\theta,\eta}\right)'\right]$, where $I_{\theta,\eta} = P\left[\dot{\ell}_{\theta,\eta}\dot{\ell}'_{\theta,\eta}\right]$. An intuitive justification for the form of the efficient score is that some information for estimating $\theta$ is lost due to a lack of knowledge about $\eta$. The amount subtracted off of the efficient score, $\Pi_{\theta,\eta}\dot{\ell}_{\theta,\eta}$, is the minimum possible amount for regular estimators when $\eta$ is unknown.

Consider again the semiparametric regression model (1.1), where $Y = \beta'Z + e$, $\mathrm{E}[e|Z] = 0$ and $\mathrm{E}[e^2|Z] \leq K < \infty$ almost surely, and where we observe $(Y, Z)$, with the joint density $\eta$ of $(e, Z)$ satisfying $\int_{\mathbb{R}} e\eta(e, Z)de = 0$ almost surely. Assume $\eta$ has partial derivative with respect to the first argument, $\dot{\eta}_1$, satisfying $\dot{\eta}_1/\eta \in L_2(P_{\beta,\eta})$, and hence $\dot{\eta}_1/\eta \in L_2^0(P_{\beta,\eta})$,

where $P_{\beta,\eta}$ is the joint distribution of $(Y, Z)$. The Euclidean parameter of interest in this semiparametric model is $\theta = \beta$. The score for $\beta$, assuming $\eta$ is known, is $\dot{\ell}_{\beta,\eta} = -Z(\dot{\eta}_1/\eta)(Y - \beta' Z, Z)$, where we use the shorthand $(f/g)(u, v) = f(u, v)/g(u, v)$ for ratios of functions.

One can show that the tangent set $\dot{\mathcal{P}}^{(\eta)}_{P_{\beta,\eta}}$ for $\eta$ is the subset of $L_2^0(P_{\beta,\eta})$ which consists of all functions $g(e, Z) \in L_2^0(P_{\beta,\eta})$ which satisfy

$$\mathrm{E}[eg(e, Z)|Z] = \frac{\int_{\mathbb{R}} eg(e, Z)\eta(e, Z)de}{\int_{\mathbb{R}} \eta(e, Z)de} = 0,$$

almost surely. One can also show that this set is the orthocomplement in $L_2^0(P_{\beta,\eta})$ of all functions of the form $ef(Z)$, where $f$ satisfies $P_{\beta,\eta} f^2(Z) < \infty$. This means that $\tilde{\ell}_{\beta,\eta} = (I - \Pi_{\beta,\eta})\dot{\ell}_{\beta,\eta}$ is the projection in $L_2^0(P_{\beta,\eta})$ of $-Z(\dot{\eta}_1/\eta)(e, Z)$ onto $\{ef(Z) : P_{\beta,\eta} f^2(Z) < \infty\}$, where $I$ is the identity. Thus

$$\tilde{\ell}_{\beta,\eta}(Y, Z) = \frac{-Ze \int_{\mathbb{R}} \dot{\eta}_1(e, Z)ede}{P_{\beta,\eta}[e^2|Z]} = -\frac{Ze(-1)}{P_{\beta,\eta}[e^2|Z]} = \frac{Z(Y - \beta' Z)}{P_{\beta,\eta}[e^2|Z]},$$

where the second-to-last step follows from the identity $\int_{\mathbb{R}} \dot{\eta}_1(e, Z)ede = \partial \int_{\mathbb{R}} \eta(te, Z)de/(\partial t)|_{t=1}$, and the last step follows since $e = Y - \beta' Z$. When the function $z \mapsto P_{\beta,\eta}[e^2|Z = z]$ is non-constant in $z$, $\tilde{\ell}_{\beta,\eta}(Y, Z)$ is not proportional to $Z(Y - \beta' Z)$, and the estimator $\hat{\beta}$ defined in Chapter 1 will not be efficient. We will discuss efficient estimation for this model in greater detail in Chapter 4.

Two very useful tools for computing efficient scores are score and information operators. Although we will provide more precise definitions in Parts II and III, operators are maps between spaces of functions. Returning to the generic semiparametric model $\{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$, sometimes it is easier to represent an element $g$ in the tangent set for $\eta$, $\dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}}$, as $B_{\theta,\eta} b$, where $b$ is an element of another set $\mathbb{H}_\eta$ and $B_{\theta,\eta}$ is an operator satisfying $\dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}} = \{B_{\theta,\eta} b : b \in \mathbb{H}_\eta\}$. Such an operator is a score operator. The adjoint of the score operator $B_{\theta,\eta} : \mathbb{H}_\eta \mapsto L_2^0(P_{\theta,\eta})$ is another operator $B^*_{\theta,\eta} : L_2^0(P_{\theta,\eta}) \mapsto \overline{\mathrm{lin}}\, \mathbb{H}_\eta$ which is similar in spirit to a transpose for matrices. Here we use $\overline{\mathrm{lin}}\, \mathbb{A}$ to denote *closed linear span* (the linear space consisting of all linear combinations) of $\mathbb{A}$. Additional details on adjoints and methods for computing them will be described in Part III. One can now define the information operator $B^*_{\theta,\eta} B_{\theta,\eta} : H_\eta \mapsto \overline{\mathrm{lin}}\, H_\eta$. If $B^*_{\theta,\eta} B_{\theta,\eta}$ has an inverse, then it can be shown that the efficient score for $\theta$ has the form $\tilde{\ell}_{\theta,\eta} = \left(I - B_{\theta,\eta}\left[B^*_{\theta,\eta} B_{\theta,\eta}\right]^{-1} B^*_{\theta,\eta}\right)\dot{\ell}_{\theta,\eta}$.

To illustrate these methods, consider the Cox model for right-censored data introduced in Chapter 1. In this setting, we observe a sample of $n$ realizations of $X = (V, d, Z)$, where $V = T \wedge C$, $d = 1\{V = T\}$, $Z \in$

$\mathbb{R}^k$ is a covariate vector, $T$ is a failure time, and $C$ is a censoring time. We assume that $T$ and $C$ are independent given $Z$, that $T$ given $Z$ has integrated hazard function $e^{\beta' Z}\Lambda(t)$ for $\beta$ in an open subset $B \subset \mathbb{R}^k$ and $\Lambda$ is continuous and monotone increasing with $\Lambda(0) = 0$, and that the censoring distribution does not depend on $\beta$ or $\Lambda$ (i.e., censoring is uninformative). Define the counting and at-risk processes $N(t) = 1\{V \le t\}d$ and $Y(t) = 1\{V \ge t\}$, and let $M(t) = N(t) - \int_0^t Y(s)e^{\beta' Z}d\Lambda(s)$. For some $0 < \tau < \infty$ with $P\{C \ge \tau\} > 0$, let $H$ be the set of all $\Lambda$'s satisfying our criteria with $\Lambda(\tau) < \infty$. Now the set of models $\mathcal{P}$ is indexed by $\beta \in B$ and $\Lambda \in H$. We let $P_{\beta,\Lambda}$ be the distribution of $(V, d, Z)$ corresponding to the given parameters.

The likelihood for a single observation is thus proportional to $p_{\beta,\Lambda}(X) = \left[e^{\beta' Z}\lambda(V)\right]^d \exp\left[-e^{\beta' Z}\Lambda(V)\right]$, where $\lambda$ is the derivative of $\Lambda$. Now let $L_2(\Lambda)$ be the set of measurable functions $b : [0, \tau] \mapsto \mathbb{R}$ with $\int_0^\tau b^2(s)d\Lambda(s) < \infty$. If $b \in L_2(\Lambda)$ is bounded, then $\Lambda_t(s) = \int_0^s e^{tb(u)}d\Lambda(u) \in H$ for all $t$. The score function $\partial \log p_{\beta+ta,\Lambda_t}(X)/(\partial t)|_{t=0}$ is thus $\int_0^\tau [a'Z + b(s)]\, dM(s)$, for any $a \in \mathbb{R}^k$. The score function for $\beta$ is therefore $\dot{\ell}_{\beta,\Lambda}(X) = ZM(\tau)$, while the score function for $\Lambda$ is $\int_0^\tau b(s)dM(s)$. In fact, one can show that there exists one-dimensional submodels $\Lambda_t$ such that $\log p_{\beta+ta,\Lambda_t}$ is differentiable with score $a'\dot{\ell}_{\beta,\Lambda}(X) + \int_0^\tau b(s)dM(s)$, for any $b \in L_2(\Lambda)$ and $a \in \mathbb{R}^k$.

The operator $B_{\beta,\Lambda} : L_2(\Lambda) \mapsto L_2^0(P_{\beta,\Lambda})$, given by $B_{\beta,\Lambda}(b) = \int_0^\tau b(s)dM(s)$, is the score operator which generates the tangent set for $\Lambda$, $\dot{\mathcal{P}}^{(\Lambda)}_{P_{\beta,\Lambda}} \equiv \{B_{\beta,\Lambda}b : b \in L_2(\Lambda)\}$. It can be shown that this tangent space spans all square-integrable score functions for $\Lambda$ generated by parametric submodels. The adjoint operator can be shown to be $B^*_{\beta,\Lambda} : L_2(P_{\beta,\Lambda}) \mapsto L_2(\Lambda)$, where $B^*_{\beta,\Lambda}(g)(t) = P_{\beta,\Lambda}[g(X)dM(t)]/d\Lambda(t)$. The information operator $B^*_{\beta,\Lambda}B_{\beta,\Lambda} : L_2(\Lambda) \mapsto L_2(\Lambda)$ is thus

$$B^*_{\beta,\Lambda}B_{\beta,\Lambda}(b)(t) = \frac{P_{\beta,\Lambda}\left[\int_0^\tau b(s)dM(s)dM(u)\right]}{d\Lambda(u)} = P_{\beta,\Lambda}\left[Y(t)e^{\beta' Z}\right] b(t),$$

using martingale methods.

Since $B^*_{\beta,\Lambda}\left(\dot{\ell}_{\beta,\Lambda}\right)(t) = P_{\beta,\Lambda}\left[ZY(t)e^{\beta' Z}\right]$, we have that the efficient score for $\beta$ is

$$\begin{aligned} \tilde{\ell}_{\beta,\Lambda} &= \left(I - B_{\beta,\Lambda}\left[B^*_{\beta,\Lambda}B_{\beta,\Lambda}\right]^{-1} B^*_{\beta,\Lambda}\right)\dot{\ell}_{\beta,\Lambda} \\ &= \int_0^\tau \left\{ Z - \frac{P_{\beta,\Lambda}\left[ZY(t)e^{\beta' Z}\right]}{P_{\beta,\Lambda}\left[Y(t)e^{\beta' Z}\right]} \right\} dM(t). \end{aligned} \tag{3.3}$$

When $\tilde{I}_{\beta,\Lambda} \equiv P_{\beta,\Lambda}\left[\tilde{\ell}_{\beta,\Lambda}\tilde{\ell}'_{\beta,\Lambda}\right]$ is positive definite, the resulting efficient influence function is $\tilde{\psi}_{\beta,\Lambda} \equiv \tilde{I}^{-1}_{\beta,\Lambda}\tilde{\ell}_{\beta,\Lambda}$. Since the estimator $\hat{\beta}_n$ obtained from

maximizing the *partial likelihood*

$$\tilde{L}_n(\beta) = \prod_{i=1}^{n} \left( \frac{e^{\beta' Z_i}}{\sum_{j=1}^{n} 1\{V_j \geq V_i\} e^{\beta' Z_j}} \right)^{d_i} \tag{3.4}$$

can be shown to satisfy $\sqrt{n}(\hat{\beta}_n - \beta) = \sqrt{n}\mathbb{P}_n \tilde{\psi}_{\beta,\Lambda} + o_P(1)$, this estimator is efficient.

Returning to our discussion of score and information operators, these operators are also useful for generating scores for the entire model, not just for the nuisance component. With semiparametric models having score functions of the form $a'\dot{\ell}_{\theta,\eta} + B_{\theta,\eta} b$, for $a \in \mathbb{R}^k$ and $b \in \mathbb{H}_\eta$, we can define a new operator $A_{\beta,\eta} : \{(a,b) : a \in \mathbb{R}^k, b \in \operatorname{lin} \mathbb{H}_\eta\} \mapsto L_2^0(P_{\theta,\eta})$ where $A_{\beta,\eta}(a,b) = a'\dot{\ell}_{\theta,\eta} + B_{\theta,\eta} b$. More generally, we can define the score operator $A_\eta : \operatorname{lin} \mathbb{H}_\eta \mapsto L_2(P_\eta)$ for the model $\{P_\eta : \eta \in H\}$, where $H$ indexes the entire model and may include both parametric and nonparametric components, and where $\operatorname{lin} \mathbb{H}_\eta$ indexes directions in $H$. Let the parameter of interest be $\psi(P_\eta) = \chi(\eta) \in \mathbb{R}^k$. We assume there exists a linear operator $\dot{\chi} : \operatorname{lin} \mathbb{H}_\eta \mapsto \mathbb{R}^k$ such that, for every $b \in \operatorname{lin} \mathbb{H}_\eta$, there exists a one-dimensional submodel $\{P_{\eta_t} : \eta_t \in H, t \in N_\epsilon\}$ satisfying

$$\int \left[ \frac{(dP_{\eta_t})^{1/2} - (dP_\eta)^{1/2}}{t} - \frac{1}{2} A_\eta b (dP_\eta)^{1/2} \right]^2 \to 0,$$

as $t \downarrow 0$, and $\partial \chi(\eta_t)/(\partial t)|_{t=0} = \dot{\chi}(b)$.

We require $\mathbb{H}_\eta$ to have a suitable *inner product* $\langle \cdot, \cdot \rangle_\eta$, where an inner product is an operation on elements in $\mathbb{H}_\eta$ with the property that $\langle a, b \rangle_\eta = \langle b, a \rangle_\eta$, $\langle a+b, c \rangle_\eta = \langle a, c \rangle_\eta + \langle b, c \rangle_\eta$, $\langle a, a \rangle_\eta \geq 0$, and $\langle a, a \rangle_\eta = 0$ if and only if $a = 0$, for all $a, b, c \in \mathbb{H}_\eta$. The efficient influence function is the solution $\tilde{\psi}_{P_\eta} \in \overline{R}(A_\eta) \subset L_2^0(P_\eta)$ of

$$A_\eta^* \tilde{\psi}_{P_\eta} = \tilde{\chi}_\eta, \tag{3.5}$$

where $R$ denotes range, $\overline{B}$ denotes closure of the set $B$, $A_\eta^*$ is the adjoint of $A_\eta$, and $\tilde{\chi}_\eta \in \mathbb{H}_\eta$ satisfies $\langle \tilde{\chi}_\eta, b \rangle_\eta = \dot{\chi}_\eta(b)$ for all $b \in \mathbb{H}_\eta$. Methods for obtaining such a $\tilde{\chi}_\eta$ will be given in Part III. When $A_\eta^* A_\eta$ is invertible, then the solution to (3.5) can be written $\tilde{\psi}_{P_\eta} = A_\eta \left( A_\eta^* A_\eta \right)^{-1} \tilde{\chi}_\eta$. In Chapter 4, we will illustrate this approach to derive efficient estimators for all parameters of the Cox model.

Returning to the semiparametric model setting, where $\mathcal{P} = \{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$, $\Theta$ is an open subset of $\mathbb{R}^k$, and $H$ is a set, the efficient score can be used to derive *estimating equations* for computing efficient estimators of $\theta$. An estimating equation is a data dependent function $\Psi_n : \Theta \mapsto \mathbb{R}^k$ for which an approximate zero yields a Z-estimator for $\theta$. When $\Psi_n(\tilde{\theta})$ has the form $\mathbb{P}_n \hat{\ell}_{\tilde{\theta},n}$, where $\hat{\ell}_{\tilde{\theta},n}(X | X_1, \ldots, X_n)$ is a function for the generic observation $X$ which depends on the value of $\tilde{\theta}$ and the sample data $X_1, \ldots, X_n$,

we have the following estimating equation result (the proof of which will be given in Part III, Page 350):

THEOREM 3.1 *Suppose that the model* $\{P_{\theta,\eta} : \theta \in \Theta\}$, *where* $\Theta \subset \mathbb{R}^k$, *is differentiable in quadratic mean with respect to* $\theta$ *at* $(\theta, \eta)$ *and let the efficient information matrix* $\tilde{I}_{\theta,\eta}$ *be nonsingular. Let* $\hat{\theta}_n$ *satisfy* $\sqrt{n}\mathbb{P}_n\hat{\ell}_{\hat{\theta}_n,n} = o_P(1)$ *and be consistent for* $\theta$. *Also assume that* $\hat{\ell}_{\hat{\theta}_n,n}$ *is contained in a* $P_{\theta,\eta}$*-Donsker class with probability tending to 1 and that the following conditions hold:*

$$P_{\hat{\theta}_n,\eta}\hat{\ell}_{\hat{\theta}_n,n} = o_P(n^{-1/2} + \|\hat{\theta}_n - \theta\|), \tag{3.6}$$

$$P_{\theta,\eta}\left\|\hat{\ell}_{\hat{\theta}_n,n} - \tilde{\ell}_{\theta,\eta}\right\|^2 \xrightarrow{\mathrm{P}} 0, \quad P_{\hat{\theta}_n,\eta}\left\|\hat{\ell}_{\hat{\theta}_n,n}\right\|^2 = O_P(1). \tag{3.7}$$

*Then* $\hat{\theta}_n$ *is asymptotically efficient at* $(\theta, \eta)$.

Returning to the Cox model example, the profile likelihood score is the partial likelihood score $\Psi_n(\tilde{\beta}) = \mathbb{P}_n\hat{\ell}_{\tilde{\beta},n}$, where

$$\hat{\ell}_{\tilde{\beta},n}(X = (V, d, Z)|X_1, \ldots, X_n) = \int_0^\tau \left\{ Z - \frac{\mathbb{P}_n\left[ZY(t)e^{\tilde{\beta}'Z}\right]}{\mathbb{P}_n\left[Y(t)e^{\tilde{\beta}'Z}\right]} \right\} dM_{\tilde{\beta}}(t), \tag{3.8}$$

and $M_{\tilde{\beta}}(t) = N(t) - \int_0^t Y(u)e^{\tilde{\beta}'Z}d\Lambda(u)$. We will show in Chapter 4 that all the conditions of Theorem 3.1 are satisfied for the root of $\Psi_n(\tilde{\beta}) = 0$, $\hat{\beta}_n$, and thus the partial likelihood yields efficient estimation of $\beta$.

Returning to the general semiparametric setting, even if an estimating equation $\Psi_n$ is not close enough to $\mathbb{P}_n\tilde{\ell}_{\theta,\eta}$ to result in an efficient estimator, frequently the estimator will still result in a $\sqrt{n}$-consistent estimator which is precise enough to be useful. In some cases, the computational effort needed to obtain an efficient estimator may be too great a cost, and one must settle for an inefficient estimating equation that works. Even in these settings, some modifications in the estimating equation can often be made which improve efficiency while maintaining computability. This issue will be explored in greater detail in Part III.

## 3.3 Maximum Likelihood Estimation

The most common approach to efficient estimation is based on modifications of maximum likelihood estimation that lead to efficient estimates. These modifications, which we will call "likelihoods," are generally not really likelihoods (products of densities) because of complications resulting

from the presence of an infinite dimensional nuisance parameter. Consider estimating an unknown real density $f(x)$ from an i.i.d. sample $X_1, \ldots, X_n$. The likelihood is $\prod_{i=1}^n f(X_i)$, and the maximizer over all densities has arbitrarily high peaks at the observations, with zero at the other values, and is therefore not a density. This problem can be fixed by using an empirical likelihood $\prod_{i=1}^n p_i$, where $p_1, \ldots, p_n$ are the masses assigned to the observations indexed by $i = 1, \ldots, n$ and are constrained to satisfy $\sum_{i=1}^n p_i = 1$. This leads to the empirical distribution function estimator, which is known to be fully efficient.

Consider again the Cox model for right-censored data explored in the previous section. The density for a single observation $X = (V, d, Z)$ is proportional to $\left[e^{\beta' Z}\lambda(V)\right]^d \exp\left[-e^{\beta' Z}\Lambda(V)\right]$. Maximizing the likelihood based on this density will result in the same problem raised in the previous paragraph. A likelihood that works is the following, which assigns mass only at observed failure times:

$$L_n(\beta, \Lambda) = \prod_{i=1}^n \left[e^{\beta' Z_i}\Delta\Lambda(V_i)\right]^{d_i} \exp\left[-e^{\beta' Z_i}\Lambda(V_i)\right], \tag{3.9}$$

where $\Delta\Lambda(t)$ is the jump size of $\Lambda$ at $t$. For each value of $\beta$, one can maximize or *profile* $L_n(\beta, \Lambda)$ over the "nuisance" parameter $\Lambda$ to obtain the profile likelihood $pL_n(\beta)$, which for the Cox model is $\exp[-\sum_{i=1}^n d_i]$ times the partial likelihood (3.4). Let $\hat{\beta}$ be the maximizer of $pL_n(\beta)$. Then the maximizer $\hat{\Lambda}$ of $L_n(\hat{\beta}, \Lambda)$ is the "Breslow estimator"

$$\hat{\Lambda}(t) = \int_0^t \frac{\mathbb{P}_n dN(s)}{\mathbb{P}_n\left[Y(s)e^{\hat{\beta}' Z}\right]}.$$

We will see in Chapter 4 that $\hat{\beta}$ and $\hat{\Lambda}$ are both efficient.

Another useful class of likelihood variants are *penalized likelihoods*. Penalized likelihoods add a penalty term in order to maintain an appropriate level of smoothness for one or more of the nuisance parameters. This method is used in the partly linear logistic regression model described in Chapter 1. Other methods of generating likelihood variants that work are possible. The basic idea is that using the likelihood principle to guide estimation of semiparametric models often leads to efficient estimators for the model components which are $\sqrt{n}$ consistent. Because of the richness of this approach to estimation, one needs to verify for each new situation that a likelihood-inspired estimator is consistent, efficient and well-behaved for moderate sample sizes. Verifying efficiency usually entails demonstrating that the estimator satisfies the efficient score equation described in the previous section.

Unfortunately, there is no guarantee that the efficient score is a derivative of the log likelihood along some submodel. A way around this problem

is to use *approximately least-favorable submodels*. This is done by finding a function $\eta_t(\theta,\eta)$ such that $\eta_0(\theta,\eta)=\eta$, for all $\theta\in\Theta$ and $\eta\in H$, where $\eta_t(\theta,\eta)\in H$ for all $t$ small enough, and such that $\tilde{\kappa}_{\theta_0,\eta_0}=\tilde{\ell}_{\theta_0,\eta_0}$, where $\tilde{\kappa}_{\theta,\eta}(x)=\partial l_{\theta+t,\eta_t(\theta,\eta)}(x)/(\partial t)|_{t=0}$, $l_{\theta,\eta}(x)$ is the log-likelihood for the observed value $x$ at the parameters $(\theta,\eta)$, and where $(\theta_0,\eta_0)$ are the true parameter values. Note that we require $\tilde{\kappa}_{\theta,\eta}=\tilde{\ell}_{\theta,\eta}$ only when $(\theta,\eta)=(\theta_0,\eta_0)$. If $(\hat{\theta}_n,\hat{\eta}_n)$ is the maximum likelihood estimator, i.e., the maximizer of $\mathbb{P}_n l_{\theta,\eta}$, then the function $t\mapsto\mathbb{P}_n l_{\hat{\theta}_n+t,\eta_t(\hat{\theta}_n,\hat{\eta}_n)}$ is maximal at $t=0$, and thus $(\hat{\theta}_n,\hat{\eta}_n)$ is a zero of $\mathbb{P}_n\tilde{\kappa}_{\tilde{\theta},\tilde{\eta}}$. Now if $\hat{\theta}_n$ and $\hat{\ell}_{\tilde{\theta},n}=\tilde{\kappa}_{\tilde{\theta},\hat{\eta}_n}$ satisfy the conditions of Theorem 3.1 at $(\theta,\eta)=(\theta_0,\eta_0)$, then the maximum likelihood estimator $\hat{\theta}_n$ is asymptotically efficient at $(\theta_0,\eta_0)$. We will explore in Part III how this flexibility is helpful for certain models.

Consider now the special case that both $\theta$ and $\eta$ are $\sqrt{n}$ consistent in the model $\{P_{\theta,\eta}:\theta\in\Theta,\eta\in H\}$, where $\Theta\subset\mathbb{R}^k$. Let $l_{\theta,\eta}$ be the log-likelihood for a single observation, and let $\hat{\theta}_n$ and $\hat{\eta}_n$ be the corresponding maximum likelihood estimators. Since $\theta$ is finite-dimensional, the log-likelihood can be varied with respect to $\theta$ in the usual way so that the maximum likelihood estimators satisfy $\mathbb{P}_n\dot{\ell}_{\hat{\theta}_n,\hat{\eta}_n}=0$.

In contrast, varying $\eta$ in the log-likelihood is more complex. We can typically use a subset of the submodels $t\mapsto\eta_t$ used for defining the tangent set and information for $\eta$ in the model. This is easiest when the score for $\eta$ is expressed as a score operator $B_{\theta,\eta}$ working on a set of indices $h\in\mathcal{H}$ (similar to what was described in Section 3.2). The likelihood equation for $\eta$ is then usually of the form $\mathbb{P}_n B_{\hat{\theta}_n,\hat{\eta}_n}h-P_{\hat{\theta}_n,\hat{\eta}_n}B_{\hat{\theta}_n,\hat{\eta}_n}h=0$ for all $h\in\mathcal{H}$. Note that we have forced the scores to be mean zero by subtracting off the mean rather than simply assuming $P_{\theta,\eta}B_{\theta,\eta}h=0$. The approach is valid if there exists some path $t\mapsto\eta_t(\theta,\eta)$, with $\eta_0(\theta,\eta)=\eta$, such that

$$B_{\theta,\eta}h(x)-P_{\theta,\eta}B_{\theta,\eta}h \quad = \quad \partial l_{\theta,\eta_t(\theta,\eta)}(x)/(\partial t)|_{t=0} \tag{3.10}$$

for all $x$ in the sample space $\mathcal{X}$. We assume (3.10) is valid for all $h\in\mathcal{H}$, where $\mathcal{H}$ is chosen so that $B_{\theta,\eta}h(x)-P_{\theta,\eta}B_{\theta,\eta}h$ is uniformly bounded on $\mathcal{H}$ for all $x\in\mathcal{X}$, $(\theta,\eta)\in\mathbb{R}^k\times H$.

We can now express $(\hat{\theta}_n,\hat{\eta}_n)$ as a Z-estimator with estimating function $\Psi_n:\mathbb{R}^k\times H\mapsto\mathbb{R}^k\times\ell^\infty(\mathcal{H})$, where $\Psi_n=(\Psi_{n1},\Psi_{n2})$, with $\Psi_{n1}(\theta,\eta)=\mathbb{P}_n\dot{\ell}_{\theta,\eta}$ and $\Psi_{n2}(\theta,\eta)=\mathbb{P}_n B_{\theta,\eta}h-P_{\theta,\eta}B_{\theta,\eta}h$, for all $h\in\mathcal{H}$. The expectation of these maps under the true parameter $(\theta_0,\eta_0)$ is the deterministic map $\Psi=(\Psi_1,\Psi_2)$, where $\Psi_1(\theta,\eta)=P_{\theta_0,\eta_0}\dot{\ell}_{\theta,\eta}$ and $\Psi_2(\theta,\eta)=P_{\theta_0,\eta_0}B_{\theta,\eta}h-P_{\theta,\eta}B_{\theta,\eta}h$, for all $h\in\mathcal{H}$. We have constructed these estimating equations so that the maximum likelihood estimators and true parameters satisfy $\Psi_n(\hat{\theta}_n,\hat{\eta}_n)=0=\Psi(\theta_0,\eta_0)$. Provided $H$ is a subset of a normed space, we can use the Z-estimator master theorem (Theorem 2.11) to obtain weak convergence of $\sqrt{n}(\hat{\theta}_n-\theta_0,\hat{\eta}_n-\eta_0)$:

COROLLARY 3.2 *Suppose that* $\dot{\ell}_{\theta,\eta}$ *and* $B_{\theta,\eta}h$, *with* $h$ *ranging over* $\mathcal{H}$ *and with* $(\theta,\eta)$ *ranging over a neighborhood of* $(\theta_0,\eta_0)$, *are contained in a* $P_{\theta_0,\eta_0}$*-Donsker class, and that both* $P_{\theta_0,\eta_0}\left\|\dot{\ell}_{\theta,\eta}-\dot{\ell}_{\theta_0,\eta_0}\right\|^2 \xrightarrow{\mathrm{P}} 0$ *and* $\sup_{h\in\mathcal{H}} P_{\theta_0,\eta_0}$ $|B_{\theta,\eta}h - B_{\theta_0,\eta_0}h|^2 \xrightarrow{\mathrm{P}} 0$, *as* $(\theta,\eta)\to(\theta_0,\eta_0)$. *Also assume that* $\Psi$ *is Fréchet-differentiable at* $(\theta_0,\eta_0)$ *with derivative* $\dot{\Psi}_0 : \mathbb{R}^k \times lin\, H \mapsto \mathbb{R}^k \times \ell^\infty(\mathcal{H})$ *that is continuously-invertible and onto its range, with inverse* $\dot{\Psi}_0^{-1} : \mathbb{R}^k \times \ell^\infty(\mathcal{H}) \mapsto \mathbb{R}^k \times lin\, H$. *Then, provided* $(\hat{\theta}_n,\hat{\eta}_n)$ *is consistent for* $(\theta_0,\eta_0)$ *and* $\Psi_n(\hat{\theta}_n,\hat{\eta}_n) = o_P(n^{-1/2})$ *(uniformly over* $\mathbb{R}^k\times\ell^\infty(\mathcal{H})$*),* $(\hat{\theta}_n,\hat{\eta}_n)$ *is efficient at* $(\theta_0,\eta_0)$ *and* $\sqrt{n}(\hat{\theta}_n-\theta_0,\hat{\eta}_n-\eta_0) \rightsquigarrow -\dot{\Psi}_0^{-1}\mathcal{Z}$, *where* $Z$ *is the Gaussian limiting distribution of* $\sqrt{n}\Psi_n(\theta_0,\eta_0)$.

The proof of this will be given later in Part III (Page 379). A function $f : U \mapsto V$ is *onto* if, for every $v \in V$, there exists a $u \in U$ with $v = f(u)$. Note that while $\mathcal{H}$ can be a subset of the tangent set used for information calculations, it must be rich enough to ensure that the inverse of $\dot{\Psi}_0$ exists.

The efficiency in Corollary 3.2 can be shown to follow from the score operator calculations given in Section 3.2, but we will postpone further details until Part III. As was done at the end of Section 3.2, the above discussion can be completely re-expressed in terms of a single parameter model $\{P_\eta : \eta \in H\}$ with a single score operator $A_\eta : \mathcal{H}_\eta \mapsto L_2(P_\eta)$, where $H$ is a richer parameter set, including, for example, both $\Theta$ and $H$ as defined in the previous paragraphs, and where $\mathcal{H}_\eta$ is similarly enriched to include the tangent sets for all subcomponents of the model.

## 3.4 Other Topics

Other topics of importance include frequentist and Bayesian methods for constructing confidence sets. We will focus primarily on frequentist approaches in this book and only briefly discuss Bayesian methods. While the bootstrap is generally valid in the setting of Corollary 3.2, it is unclear that this remains true when the nuisance parameter converges at a rate slower than $\sqrt{n}$, even if interest is limited to the parametric component. Even when the bootstrap is valid, it may be excessively cumbersome to re-estimate the entire model for many bootstrapped data sets. We will explore this issue in more detail in Part III. We now present one approach for hypothesis testing and variance estimation for the parametric component $\theta$ of the semiparametric model $\{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$. This approach is often valid even when the nuisance parameter is not $\sqrt{n}$ consistent.

Murphy and van der Vaart demonstrated that under reasonable regularity conditions, the log-profile likelihood, $pl_n(\theta)$, (profiling over the nuisance parameter) admits the following expansion about the maximum likelihood

estimator for the parametric component $\hat{\theta}_n$:

$$\begin{aligned}(3.11)\quad \log pl_n(\tilde{\theta}_n) &= \log pl_n(\hat{\theta}_n) - \frac{1}{2} n(\tilde{\theta}_n - \hat{\theta}_n)' \tilde{I}_{\theta_0,\eta_0}(\tilde{\theta}_n - \hat{\theta}_n) \\ &\quad + o_P(\sqrt{n}\|\tilde{\theta}_n - \theta_0\| + 1)^2,\end{aligned}$$

for any estimator $\tilde{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$ (Murphy and van der Vaart, 2000). This can be shown to lead naturally to chi-square tests of full versus reduced models. Furthermore, this result demonstrates that the curvature of the log-partial likelihood can serve as a consistent estimator of the efficient information for $\theta$ at $\theta_0$, $\tilde{I}_{\theta_0,\eta_0}$, and thereby permit estimation of the limiting variance of $\sqrt{n}(\hat{\theta}_n - \theta_0)$. We will discuss this method, along with several other methods of inference, in greater detail toward the end of Part III.

## 3.5 Exercises

**3.5.1.** Consider $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where $\Theta$ is an open subset of $\mathbb{R}^k$. Assume that $\mathcal{P}$ is dominated by $\mu$, and that $\dot{\ell}_\theta(x) = \partial \log p_\theta(x)/(\partial\theta)$ exists with $P_\theta[\dot{\ell}_\theta \dot{\ell}_\theta']$ bounded and $P_\theta \|\dot{\ell}_{\tilde{\theta}} - \dot{\ell}_\theta\|^2 \to 0$ as $\tilde{\theta} \to \theta$. For each $h \in \mathbb{R}^k$, let $\epsilon > 0$ be small enough so that $\{P_t : t \in N_\epsilon\} \subset \mathcal{P}$, where for each $t \in N_\epsilon$, $P_t = P_{\theta+th}$. Show that each of these one-dimensional submodels satisfy (3.1), with $g = h'\dot{\ell}_\theta$.

**3.5.2.** Consider the paragraphs leading up to the efficient score for $\beta$ in the Cox model, given in Expression (3.3). Show that the tangent set for the nuisance parameter $\Lambda$, $\dot{\mathcal{P}}^{(\Lambda)}_{P_{\beta,\Lambda}}$, spans all square-integrable score functions for $\Lambda$ generated by parametric submodels (where the parameters for $\Lambda$ are independent of $\beta$).

**3.5.3.** Let $L_n(\beta, \Lambda)$ be the Cox model likelihood given in (3.9). Show that the corresponding profile likelihood for $\beta$, $pL_n(\beta)$, obtained by maximizing over $\Lambda$, equals $\exp[-\sum_{i=1}^n d_i]$ times the partial likelihood (3.4).

## 3.6 Notes

The linear regression example was partially inspired by Example 25.28 of van der Vaart (1998), and Theorem 3.1 is his Theorem 25.54.

# 4
# Case Studies I

We now expand upon several examples introduced in Chapters 1–3 to more fully illustrate the methods and theory we have outlined thus far. Certain technical aspects which involve concepts introduced later in the book will be glossed over to avoid getting bogged down with details. The main objective of this chapter is to initiate an appreciation for what empirical processes and efficiency calculations can accomplish.

The first example is linear regression with either mean zero or median zero residuals. In addition to efficiency calculations for model parameters, empirical processes are needed for inference on the distribution of the residuals. The second example is counting process regression for both general counting processes and the Cox model for right-censored failure time data. Empirical processes will be needed for parameter inference, and efficiency will be established under the Cox proportional hazards model for maximum likelihood estimation of both the regression parameters and the baseline hazard. The third example is the Kaplan-Meier estimator of the survival function for right-censored failure time data. Since weak convergence of the Kaplan-Meier has already been established in Chapter 2 using empirical processes, the focus in this chapter will be on efficiency calculations. The fourth example considers estimating equations for general regression models when the residual variance may be a function of the covariates. Estimation of the variance function is needed for efficient estimation. We also consider optimality of a certain class of estimating equations. The general results are illustrated with both simple linear regression and a Poisson mixture regression model. In the latter case, the mixture distribution is not $\sqrt{n}$ consistent in the uniform norm, the proof of which fact we omit. The

fifth, and final, example is partly linear logistic regression. The emphasis will be on efficient estimation of the parametric regression parameter. For the last two examples, both empirical processes and efficiency calculations will be needed.

## 4.1 Linear Regression

The semiparametric linear regression model is $Y = \beta' Z + e$, where we observe $X = (Y, Z)$ and assume $\mathrm{E}\|Z\|^2 < \infty$, $\mathrm{E}[e|Z] = 0$ and $\mathrm{E}[e^2|Z] \leq K < \infty$ almost surely, $Z$ includes the constant 1, and $\mathrm{E}[ZZ']$ is full rank. The model for $X$ is $\{P_{\beta,\eta} : \beta \in \mathbb{R}^k, \eta \in H\}$, where $\eta$ is the joint density of the residual $e$ and covariate $Z$ with partial derivative with respect to the first argument $\dot{\eta}_1$ which we assume satisfies $\dot{\eta}/\eta \in L_2(P_{\beta,\eta})$ and hence $\dot{\eta}/\eta \in L_2^0(P_{\beta,\eta})$. We consider this model under two assumptions on $\eta$: the first is that the residuals have conditional mean zero, i.e., $\int_{\mathbb{R}} u\eta(u, Z)du = 0$ almost surely; and the second is that the residuals have median zero, i.e., $\int_{\mathbb{R}} \mathrm{sign}(u)\eta(u, Z)du = 0$ almost surely.

### *4.1.1 Mean Zero Residuals*

We have already demonstrated in Section 3.2 that the usual least squares estimator $\hat{\beta} = [\mathbb{P}_n ZZ']^{-1} \mathbb{P}_n ZY$ is $\sqrt{n}$ consistent but not always efficient for $\beta$ when the only assumption we are willing to make is that the residuals have mean zero conditional on the covariates. The basic argument for this was taken from the form of the efficient score $\tilde{\ell}_{\beta,\eta}(Y, Z) = Z(Y - \beta' Z)/P_{\beta,\eta}[e^2|Z]$ which yields a distinctly different estimator than $\hat{\beta}$ when $z \mapsto P_{\beta,\eta}[e^2|Z = z]$ is non-constant in $z$. In Section 4.4 of this chapter, we will describe a data-driven procedure for efficient estimation in this context based on approximating the efficient score.

For now, however, we turn our attention to efficient estimation when we also assume that the covariates are independent of the residual. Accordingly, we denote $\eta$ to be the density of the residual $e$ and $\dot{\eta}$ to be the derivative of $\eta$. We discussed this model in Chapter 1 and pointed at that $\hat{\beta}$ is still not efficient in this setting. We also claimed that an empirical estimator $\hat{F}$ of the residual distribution, based on the residuals $Y_1 - \hat{\beta}' Z_1, \ldots, Y_n - \hat{\beta}' Z_n$, had the property that $\sqrt{n}(\hat{F} - F)$ converged in a uniform sense to a certain Guassian quantity. We now derive the efficient score for estimating $\beta$ for this model and sketch a proof of the claimed convergence of $\sqrt{n}(\hat{F} - F)$. We assume that $\eta$ is continuously differentiable with $P_{\beta,\eta}(\dot{\eta}/\eta)^2 < \infty$.

Using techniques described in Section 3.2, it is not hard to verify that the tangent set for the full model is the linear span of $-(\dot{\eta}/\eta)(e)a' Z + b(e)$, as $a$ spans $\mathbb{R}^k$ and $b$ spans the tangent set $\dot{\mathcal{P}}^{(\eta)}_{P_{\beta,\eta}}$ for $\eta$, consisting of all functions in $L_2^0(e)$ which are orthogonal to $e$, where we use $L_2^0(U)$ to denote all mean

zero real functions $f$ of the random variable $U$ for which $Pf^2(U) < \infty$. The structure of the tangent set follows from the constraint that $\int_{\mathbb{R}} e\eta(e)de = 0$.

The projection of the usual score for $\beta$, $\dot{\ell}_{\beta,\eta} \equiv -(\dot{\eta}/\eta)(e)Z$, onto $\dot{\mathcal{P}}^{(\eta)}_{P_{\beta,\eta}}$ is a function $h \in \dot{\mathcal{P}}^{(\eta)}_{P_{\beta,\eta}}$ such that $-(\dot{\eta}/\eta)(e)Z - h(e)$ is uncorrelated with $b(e)$ for all $b \in \dot{\mathcal{P}}^{(\eta)}_{P_{\beta,\eta}}$. We will now verify that $h(e) = -(\dot{\eta}/\eta)(e)\mu - e\mu/\sigma^2$, where $\mu \equiv \mathrm{E}[Z]$ and $\sigma^2 = \mathrm{E}[e^2]$. It easy to see that $h$ is square integrable. Moreover, since $\int_{\mathbb{R}} e\dot{\eta}(e)de = -1$, $h$ has mean zero. Thus $h \in \dot{\mathcal{P}}^{(\eta)}_{P_{\beta,\eta}}$. It also follows that $-(\dot{\eta}/\eta)(e)Z - h(e)$ is orthogonal to $\dot{\mathcal{P}}^{(\eta)}_{P_{\beta,\eta}}$, after noting that $-(\dot{\eta}/\eta)(e)Z - h(e) = -(\dot{\eta}/\eta)(e)(Z-\mu) + e\mu/\sigma^2 \equiv \tilde{\ell}_{\beta,\eta}$ is orthogonal to all square-integrable mean zero functions $b(e)$ which satisfy $P_{\beta,\eta}b(e)e = 0$.

Thus the efficient information is

$$(4.1)\ \tilde{I}_{\beta,\eta} \equiv P_{\beta,\eta}\left[\tilde{\ell}_{\beta,\eta}\tilde{\ell}'_{\beta,\eta}\right] = P_{\beta,\eta}\left[\left\{\frac{\dot{\eta}}{\eta}(e)\right\}^2 (Z-\mu)(Z-\mu)'\right] + \frac{\mu\mu'}{\sigma^2}.$$

Since

$$1 = \left(\int_{\mathbb{R}} e\dot{\eta}(e)de\right)^2 = \left(\int_{\mathbb{R}} e\frac{\dot{\eta}}{\eta}(e)\eta(e)de\right)^2 \leq \sigma^2 P_{\beta\eta}\left\{\frac{\dot{\eta}}{\eta}(e)\right\}^2,$$

we have that

$$(4.1) \geq P_{\beta,\eta}\left[\frac{(Z-\mu)(Z-\mu)'}{\sigma^2}\right] + \frac{\mu\mu'}{\sigma^2} = \frac{P[ZZ']}{\sigma^2},$$

where, for two $k \times k$ symmetric matrices $A$ and $B$, $A \geq B$ means that $A - B$ is positive semidefinite. Thus the efficient estimator can have strictly lower variance than the least-squares estimator $\hat{\beta}$.

Developing a procedure for calculating an asymptotically efficient estimator appears to require nonparametric estimation of $\dot{\eta}/\eta$. We will show in the semiparametric inference case studies of Chapter 22 how to accomplish this via the residual distribution estimator $\hat{F}$ defined above. We now turn our attention to verifying that $\sqrt{n}(\hat{F} - F)$ converges weakly to a tight, mean zero Gaussian process. This result can also be used to check normality of the residuals. Recall that if the residuals are Gaussian, the least-squares estimator is fully efficient.

Now, using $P = P_{\beta,\eta}$,

$$\begin{aligned}
\sqrt{n}(\hat{F}(v) - F(v)) &= \sqrt{n}\left[\mathbb{P}_n 1\{Y - \hat{\beta}'Z \leq v\} - P1\{Y - \beta'Z \leq v\}\right] \\
&= \sqrt{n}(\mathbb{P}_n - P)1\{Y - \hat{\beta}'Z \leq v\} \\
&\quad + \sqrt{n}P\left[1\{Y - \hat{\beta}'Z \leq v\} - 1\{Y - \beta'Z \leq v\}\right] \\
&= U_n(v) + V_n(v).
\end{aligned}$$

We will show in Part II that $\left\{1\{Y - b'Z \leq v\} : b \in \mathbb{R}^k, v \in \mathbb{R}\right\}$ is a VC (and hence Donsker) class of functions. Thus, since

$$\sup_{v \in \mathbb{R}} P\left[1\{Y - \hat{\beta}'Z \leq v\} - 1\{Y - \beta'Z \leq v\}\right]^2 \xrightarrow{\mathrm{P}} 0,$$

we have that $U_n(v) = \sqrt{n}(\mathbb{P}_n - P)1\{Y - \beta'Z \leq v\} + \epsilon_n(v)$, where $\sup_{v \in \mathbb{R}} |\epsilon_n(v)| \xrightarrow{\mathrm{P}} 0$. It is not difficult to show that

$$V_n(v) = P\left[\int_v^{v+(\hat{\beta}-\beta)'Z} \eta(u)du\right]. \tag{4.2}$$

We leave it as an exercise (see Exercise 4.6.1) to show that for any $u, v \in \mathbb{R}$,

$$|\eta(u) - \eta(v)| \leq |F(u) - F(v)|^{1/2} \left(P\left\{\frac{\dot{\eta}}{\eta}(e)\right\}^2\right)^{1/2}. \tag{4.3}$$

Thus $\eta$ is both bounded and equicontinuous, and thus by (4.2), $V_n(v) = \sqrt{n}(\hat{\beta} - \beta)'\mu\eta(v) + \tilde{\epsilon}_n(v)$, where $\sup_{v \in \mathbb{R}} |\tilde{\epsilon}_n(v)| \xrightarrow{\mathrm{P}} 0$. Hence $\hat{F}$ is asymptotically linear with influence function

$$\check{\psi}(v) = 1\{e \leq v\} - F(v) + eZ' \left\{P[ZZ']\right\}^{-1} \mu\eta(v),$$

and thus, since this influence function is a Donsker class (as the sum of two obviously Donsker classes), $\sqrt{n}(\hat{F} - F)$ converges weakly to a tight, mean zero Gaussian process.

### *4.1.2 Median Zero Residuals*

We have already established in Section 2.2.6 that when the residuals have median zero and are independent of the covariates, then the least-absolute-deviation estimator $\hat{\beta} \equiv \mathrm{argmin}_{b \in \mathbb{R}^k} \mathbb{P}_n|Y - b'Z|$ is consistent for $\beta$ and $\sqrt{n}(\hat{\beta} - \beta)$ is asymptotically linear with influence function

$$\check{\psi} = \left\{2\eta(0)P[ZZ']\right\}^{-1} Z\mathrm{sign}(e).$$

Thus $\sqrt{n}(\hat{\beta} - \beta)$ is asymptotically mean zero Gaussian with covariance equal to $\left\{4\eta^2(0)P[ZZ']\right\}^{-1}$. In this section, we will study efficiency for this model and show that $\hat{\beta}$ is not in general fully efficient. Before doing this, however, we will briefly study efficiency in the the more general model where we only assume $\mathrm{E}[\mathrm{sign}(e)|Z] = 0$ almost surely.

Under this more general model, the joint density of $(e, Z)$, $\eta$, must satisfy $\int_{\mathbb{R}} \mathrm{sign}(e)\eta(e, Z)de = 0$ almost surely. As we did when we studied the conditionally mean zero residual case in Section 3.2, assume $\eta$ has partial

derivative with respect to the first argument, $\dot{\eta}_1$, satisfying $\dot{\eta}_1/\eta \in L_2(P_{\beta,\eta})$. Clearly, $(\dot{\eta}_1/\eta)(e, Z)$ also has mean zero. The score for $\beta$, assuming $\eta$ is known, is $\dot{\ell}_{\beta,\eta} = -Z(\dot{\eta}_1/\eta)(Y - \beta' Z, Z)$.

Similar to what was done in Section 3.2 for the conditionally mean zero case, it can be shown that the tangent set $\dot{\mathcal{P}}^{(\eta)}_{P_{\beta,\eta}}$ for $\eta$ is the subset of $L_2^0(P_{\beta,\eta})$ which consists of all functions $g(e, Z) \in L_2^0(P_{\beta,\eta})$ which satisfy

$$\mathrm{E}[\mathrm{sign}(e)g(e, Z)|Z] = \frac{\int_{\mathbb{R}} \mathrm{sign}(e)g(e, Z)\eta(e, Z)de}{\int_{\mathbb{R}} \eta(e, Z)de} = 0,$$

almost surely. It can also be shown that this set is the orthocomplement in $L_2^0(P_{\beta,\eta})$ of all functions of the form $\mathrm{sign}(e)f(Z)$, where $f$ satisfies $P_{\beta,\eta}f^2(Z) < \infty$. Hence the efficient score $\tilde{\ell}_{\beta,\eta}$ is the projection in $L_2^0(P_{\beta,\eta})$ of $-Z(\dot{\eta}_1/\eta)(e, Z)$ onto $\{\mathrm{sign}(e)f(Z) : P_{\beta,\eta}f^2(Z) < \infty\}$. Hence

$$\tilde{\ell}_{\beta,\eta}(Y, Z) = -Z\mathrm{sign}(e) \int_{\mathbb{R}} \dot{\eta}_1(e, Z)\mathrm{sign}(e)de = 2Z\mathrm{sign}(e)\eta(0, Z),$$

where the second equality follows from the facts that $\int_{\mathbb{R}} \dot{\eta}_1(e, Z)de = 0$ and $\int_{-\infty}^{0} \dot{\eta}_1(e, Z)de = \eta(0, Z)$. When $\eta(0, z)$ is non-constant in $z$, $\tilde{\ell}_{\beta,\eta}(Y, Z)$ is not proportional to $\mathrm{sign}(Z)(Y - \beta' Z)$, and thus the least-absolute-deviation estimator is not efficient in this instance. Efficient estimation in this situation appears to require estimation of $\eta(0, Z)$, but we will not pursue this further.

We now return our attention to the setting where the median zero residuals are independent of the covariates. We will also assume that $0 < \eta(0) < \infty$. Recall that in this setting, the usual score for $\beta$ is $\dot{\ell}_{\beta,\eta} \equiv -Z(\dot{\eta}/\eta)(e)$, where $\dot{\eta}$ is the derivative of the density $\eta$ of $e$. If we temporarily make the fairly strong assumption that the residuals have a Laplace density (i.e., $\eta(e) = (\nu/2)\exp(-\nu|e|)$ for a real parameter $\nu > 0$), then the usual score simplifies to $Z\mathrm{sign}(e)$, and thus the least-absolute deviation estimator is fully efficient for this special case.

Relaxing the assumptions to allow for arbitrary median zero density, we can follow arguments similar to those we used in the previous section to obtain that the tangent set for $\eta$, $\dot{\mathcal{P}}^{(\eta)}_{P_{\beta,\eta}}$, consists of all functions in $L_2^0(e)$ which are orthogonal to $\mathrm{sign}(e)$. The structure of the tangent set follows from the median zero residual constraint which can be expressed as $\int_{\mathbb{R}} \mathrm{sign}(e)\eta(e)de = 0$. The projection of $\dot{\ell}_{\beta,\eta}$ on $\dot{\mathcal{P}}^{(\eta)}_{\beta,\eta}$ is a function $h \in \dot{\mathcal{P}}^{(\eta)}_{P_{\beta,\eta}}$ such that $-(\dot{\eta}/\eta)(e)Z - h(e)$ is uncorrelated with $b(e)$ for all $b \in \dot{\mathcal{P}}^{(\eta)}_{P_{\beta,\eta}}$.

We will now prove that $h(e) = -(\dot{\eta}/\eta)(e)\mu - 2\eta(0)\mathrm{sign}(e)\mu$ satisfies the above constraints. First, it is easy to see that $h(e)$ is square-integrable. Second, since $-\int_{\mathbb{R}} \mathrm{sign}(e)\dot{\eta}(e)de = 2\eta(0)$, we have that $\mathrm{sign}(e)h(e)$ has zero expectation. Thus $h \in \dot{\mathcal{P}}^{(\eta)}_{P_{\beta,\eta}}$. Thirdly, it is straightforward to verify that $\dot{\ell}_{\beta,\eta}(Y, Z) - h(e) = -(\dot{\eta}/\eta)(e)(Z - \mu) + 2\eta(0)\mathrm{sign}(e)\mu$ is orthogonal to

all elements of $\dot{\mathcal{P}}^{(\eta)}_{P_{\beta,\eta}}$. Hence the efficient score is $\tilde{\ell}_{\beta,\eta} = \dot{\ell}_{\beta,\eta} - h$. Thus the efficient information is

$$\tilde{I}_{\beta,\eta} = P_{\beta,\eta}\left[\tilde{\ell}_{\beta,\eta}\tilde{\ell}'_{\beta,\eta}\right] = P_{\beta,\eta}\left[\left\{\frac{\dot{\eta}}{\eta}(e)\right\}^2 (Z-\mu)(Z-\mu)'\right] + 4\eta^2(0)\mu\mu'.$$

Note that

$$\begin{aligned} 4\eta^2(0) &= 4\left[\int_{-\infty}^{0} \dot{\eta}(e)de\right]^2 \\ &= 4\left[\int_{-\infty}^{0} \frac{\dot{\eta}}{\eta}(e)\eta(e)de\right]^2 \\ &\leq 4\int_{-\infty}^{0}\left\{\frac{\dot{\eta}}{\eta}(e)\right\}^2 \eta(e)de \int_{-\infty}^{0}\eta(e)de \\ &= 2\int_{-\infty}^{0}\left\{\frac{\dot{\eta}}{\eta}(e)\right\}^2 \eta(e)de. \end{aligned} \tag{4.4}$$

Similar arguments yield that

$$4\eta^2(0) \leq 2\int_{0}^{\infty}\left\{\frac{\dot{\eta}}{\eta}(e)\right\}^2 \eta(e)de.$$

Combining this last inequality with (4.4), we obtain that

$$4\eta^2(0) \leq \int_{\mathbb{R}}\left\{\frac{\dot{\eta}}{\eta}(e)\right\}^2 \eta(e)de.$$

Hence the efficient estimator can have strictly lower variance than the least-absolute-deviation estimator in this situation.

While the least-absolute-deviation estimator is only guaranteed to be fully efficient for Laplace distributed residuals, it does have excellent robustness properties. In particular, the *breakdown point* of this estimator is 50%. The breakdown point is the maximum proportion of contamination—from arbitrarily large symmetrically distributed residuals—tolerated by the estimator without resulting in inconsistency.

## 4.2 Counting Process Regression

We now examine in detail the counting process regression model considered in Chapter 1. The observed data are $X = (N, Y, Z)$, where for $t \in [0, \tau]$, $N(t)$ is a counting process and $Y(t) = 1\{V \geq t\}$ is an at-risk process based on a random time $V \geq 0$ which may depend on $N$, with $PY(0) = 1$, $\inf_Z P[Y(\tau)|Z] > 0$, $PN^2(\tau) < \infty$, and where $Z \in \mathbb{R}^k$ is a regression

covariate. The regression model (1.3) is assumed, implying $\mathrm{E}\{dN(t)|Z\} = \mathrm{E}\{Y(t)|Z\}e^{\beta' Z}d\Lambda(t)$, for some $\beta \in B \subset \mathbb{R}^k$ and continuous nondecreasing function $\Lambda(t)$ with $\Lambda(0) = 0$ and $0 < \Lambda(\tau) < \infty$. We assume $Z$ is restricted to a bounded set, $\mathrm{var}(Z)$ is positive definite, and that $B$ is open, convex and bounded. We first consider inference for $\beta$ and $\Lambda$ in this general model, and then examine the specialization to the Cox proportional hazards model for right-censored failure time data.

### *4.2.1 The General Case*

As described in Chapter 1, we estimate $\beta$ with the estimating equation $U_n(t,\beta) = \mathbb{P}_n \int_0^t [Z - E_n(s,\beta)]\, dN(s)$, where

$$E_n(t,\beta) = \frac{\mathbb{P}_n ZY(t)e^{\beta' Z}}{\mathbb{P}_n Y(t)e^{\beta' Z}}.$$

Specifically, the estimator $\hat{\beta}$ is a root of $U_n(\tau,\beta) = 0$. The estimator for $\Lambda$, is $\hat{\Lambda}(t) = \int_0^t \left[\mathbb{P}_n Y(s)e^{\hat{\beta}' Z}\right]^{-1} \mathbb{P}_n dN(s)$. We first show that $\hat{\beta}$ is consistent for $\beta$, and that $\sqrt{n}(\hat{\beta} - \beta)$ is asymptotically normal with a consistently estimable covariance matrix. We then derive consistency and weak convergence results for $\hat{\Lambda}$ and suggest a simple method of inference based on the influence function.

We first argue that certain classes of functions are Donsker, and therefore also Glivenko-Cantelli. We first present the following lemma:

LEMMA 4.1 *For $-\infty < a < b < \infty$, Let $\{X(t), t \in [a,b]\}$ be a monotone cadlag or caglad stochastic process with $P\left[|X(a)| \vee |X(b)|\right]^2 < \infty$. Then $X$ is $P$-Donsker.*

The proof will be given later in Part II. Note that we usually speak of classes of functions as being Donsker or Glivenko-Cantelli, but in Lemma 4.1, we are saying this about a process. Let $\mathcal{X}$ be the sample space for the stochastic process $\{X(t) : t \in T\}$. Then $\sqrt{n}(\mathbb{P}_n - P)X$ converges weakly in $\ell^\infty(T)$ to a tight, mean zero Gaussian process if and only if $\mathcal{F} = \{f_t : t \in T\}$ is $P$-Donsker, where for any $x \in \mathcal{X}$ and $t \in T$, $f_t(x) = x(t)$. Viewed in this manner, this modified use of the term Donsker is, in fact, not a modification at all.

Since $Y$ and $N$ both satisfy the conditions of Lemma 4.1, they are both Donsker as processes in $\ell^\infty([0,\tau])$. Trivially, the classes $\{\beta \in B\}$ and $\{Z\}$ are both Donsker classes, and therefore so is $\{\beta' Z : \beta \in B\}$ since products of bounded Donsker classes are Donsker. Now the class $\{e^{\beta' Z} : \beta \in B\}$ is Donsker since exponentiation is Lipschitz continuous on compacts. Hence $\{Y(t)e^{\beta' Z} : t \in [0,\tau], \beta \in B\}$, $\{ZY(t)e^{\beta' Z} : t \in [0,\tau], \beta \in B\}$, and $\{ZZ'Y(t)e^{\beta' Z} : t \in [0,\tau], \beta \in B\}$, are all Donsker since they are all products of bounded Donsker classes.

Now the derivative of $U_n(\tau,\beta)$ with respect to $\beta$ can be shown to be $-V_n(\beta)$, where

$$V_n(\beta)=\int_0^\tau\left[\frac{\mathbb{P}_nZZ'Y(t)e^{\beta'Z}}{\mathbb{P}_nY(t)e^{\beta'Z}}-\left\{\frac{\mathbb{P}_nZY(t)e^{\beta'Z}}{\mathbb{P}_nY(t)e^{\beta'Z}}\right\}^{\otimes2}\right]\mathbb{P}_ndN(t),$$

and where superscript $\otimes2$ denotes outer product. Because all of the classes involved are Glivenko-Cantelli and the limiting values of the denominators are bounded away from zero, we have the $\sup_{\beta\in B}|V_n(\beta)-V(\beta)|\overset{\text{as*}}{\to}0$, where

$$\begin{aligned}V(\beta) &= \int_0^\tau\left[\frac{PZZ'Y(t)e^{\beta'Z}}{PY(t)e^{\beta'Z}}-\left\{\frac{PZY(t)e^{\beta'Z}}{PY(t)e^{\beta'Z}}\right\}^{\otimes2}\right]\\ &\quad\times P\left[Y(t)e^{\beta'Z}\right]d\Lambda(t).\end{aligned}\tag{4.5}$$

After some work, it can be shown that there exists a $c>0$ such that $V(\beta)\geq c\,\mathrm{var}(Z)$, where for two symmetric matrices $A_1,A_2$, $A_1\geq A_2$ means that $A_1-A_2$ is positive semidefinite. Thus $U_n(\tau,\beta)$ is almost surely convex for all $n\geq1$ large enough. Thus $\hat\beta$ is almost surely consistent for the true parameter $\beta_0$.

Using algebra, $U_n(\tau,\beta)=\mathbb{P}_n\int_0^\tau[Z-E_n(s,\beta)]\,dM_\beta(s)$, where $M_\beta(t)=N(t)-\int_0^tY(s)e^{\beta'Z}d\Lambda_0(s)$ and $\Lambda_0$ is the true value of $\Lambda$. Let $U(t,\beta)=P\left\{\int_0^t[Z-E(s,\beta)]\,dM_\beta(s)\right\}$, where $E(t,\beta)=P\left[ZY(t)e^{\beta'Z}\right]/P\left[Y(t)e^{\beta'Z}\right]$. It is not difficult to verify that

$$\sqrt{n}\left[U_n(\tau,\hat\beta)-U(\tau,\hat\beta)\right]-\sqrt{n}\left[U_n(\tau,\beta_0)-U(\tau,\beta_0)\right]\overset{\text{P}}{\to}0,$$

since

$$\begin{aligned}&\sqrt{n}\mathbb{P}_n\int_0^\tau\left[E_n(s,\hat\beta)-E(s,\hat\beta)\right]dM_{\hat\beta}(s)\\ &=\sqrt{n}\mathbb{P}_n\int_0^\tau\left[E_n(s,\hat\beta)-E(s,\hat\beta)\right]\\ &\quad\times\left\{dM_{\beta_0}(s)-Y(s)\left[e^{\hat\beta'Z}-e^{\beta_0'Z}\right]d\Lambda_0(s)\right\}\\ &\overset{\text{P}}{\to}0.\end{aligned}\tag{4.6}$$

This follows from the following lemma (which we prove later in Part II), with $[a,b]=[0,\tau]$, $A_n(t)=E_n(t,\hat\beta)$, and $B_n(t)=\sqrt{n}\mathbb{P}_nM_{\beta_0}(t)$:

LEMMA 4.2 *Let $B_n\in D[a,b]$ and $A_n\in\ell^\infty([a,b])$ be either cadlag or caglad, and assume $\sup_{t\in(a,b]}|A_n(t)|\overset{\text{P}}{\to}0$, $A_n$ has uniformly bounded total variation, and $B_n$ converges weakly to a tight, mean zero process with sample paths in $D[a,b]$. Then $\int_a^bA_n(s)dB_n(s)\overset{\text{P}}{\to}0$.*

Thus the conditions of the Z-estimator master theorem, Theorem 2.11, are all satisfied, and $\sqrt{n}(\hat{\beta} - \beta_0)$ converges weakly to a mean zero random vector with covariance $C = V^{-1}(\beta_0)P\left[\int_0^\tau [Z - E(s,\beta_0)]\, dM_{\beta_0}(s)\right]^{\otimes 2} V^{-1}(\beta_0)$. With a little additional work, it can be verified that $C$ can be consistently estimated with $\hat{C} = V_n^{-1}(\hat{\beta})\mathbb{P}_n\left\{\int_0^\tau \left[Z - E_n(s,\hat{\beta})\right] d\hat{M}(s)\right\}^{\otimes 2} V_n^{-1}(\hat{\beta})$, where $\hat{M}(t) = N(t) - \int_0^t Y(s)e^{\hat{\beta}'Z}d\hat{\Lambda}(s)$, and where

$$\hat{\Lambda}(t) = \int_0^t \frac{\mathbb{P}_n dN(s)}{\mathbb{P}_n Y(s)e^{\hat{\beta}'Z}},$$

as defined in Chapter 1.

Let $\Lambda_0$ be the true value of $\Lambda$. Then

$$\begin{aligned}
\hat{\Lambda}(t) - \Lambda_0(t) &= \int_0^t 1\{\mathbb{P}_n Y(s) > 0\}\left\{\frac{(\mathbb{P}_n - P)dN(s)}{\mathbb{P}_n Y(s)e^{\hat{\beta}'Z}}\right\} \\
&\quad - \int_0^t 1\{\mathbb{P}_n Y(s) = 0\}\left\{\frac{PdN(s)}{\mathbb{P}_n Y(s)e^{\hat{\beta}'Z}}\right\} \\
&\quad - \int_0^t \frac{(\mathbb{P}_n - P)Y(s)e^{\hat{\beta}'Z}}{PY(s)e^{\hat{\beta}'Z}}\left\{\frac{PdN(s)}{\mathbb{P}_n Y(s)e^{\hat{\beta}'Z}}\right\} \\
&\quad - \int_0^t \frac{P\left[Y(s)\left(e^{\hat{\beta}'Z} - e^{\beta_0'Z}\right)\right]}{PY(s)e^{\hat{\beta}'Z}} d\Lambda_0(s) \\
&= A_n(t) - B_n(t) - C_n(t) - D_n(t).
\end{aligned}$$

By the smoothness of these functions of $\hat{\beta}$ and the almost sure consistency of $\hat{\beta}$, each of the processes $A_n$, $B_n$, $C_n$ and $D_n$ converge uniformly to zero, and thus $\sup_{t\in[0,\tau]}\left|\hat{\Lambda}(t) - \Lambda_0(t)\right| \overset{\text{as}*}{\to} 0$. Since $P\{\mathbb{P}_n Y(t) = 0\} \leq [P\{V < \tau\}]^n$, $B_n(t) = o_P(n^{-1/2})$, where the $o_P(n^{-1/2})$ term is uniform in $t$. It is also not hard to verify that $A_n(t) = (\mathbb{P}_n - P)\tilde{A}(t) + o_P(n^{-1/2})$, $C_n(t) = (\mathbb{P}_n - P)\tilde{C}(t) + o_P(n^{-1/2})$, where $\tilde{A}(t) = \int_0^t \left[PY(s)e^{\beta_0'Z}\right]^{-1} dN(s)$, $\tilde{C}(t) = \int_0^t \left[PY(s)e^{\beta_0'Z}\right]^{-1} Y(s)e^{\beta_0'Z}d\Lambda_0(s)$, and both remainder terms are uniform in $t$. In addition,

$$D_n(t) = (\hat{\beta} - \beta_0)' \int_0^t \left\{\frac{PZY(s)e^{\beta_0'Z}}{PY(s)e^{\beta_0'Z}}\right\} d\Lambda_0(t) + o_P(n^{-1/2}),$$

where the remainder term is again uniform in $t$.

Taking this all together, we obtain the expansion

$$\begin{aligned}\sqrt{n}\left[\hat{\Lambda}(t)-\Lambda_0(t)\right] &= \sqrt{n}(\mathbb{P}_n-P)\int_0^t \frac{dM_{\beta_0}(s)}{PY(s)e^{\beta_0' Z}} \\ &\quad -\sqrt{n}(\hat{\beta}-\beta_0)'\int_0^t E(s,\beta_0)d\Lambda_0(s)+o_P(1),\end{aligned}$$

where the remainder term is uniform in $t$. By previous arguments, $\sqrt{n}(\hat{\beta}-\beta_0)=V^{-1}(\beta_0)\mathbb{P}_n\int_0^\tau [Z-E(s,\beta_0)]\,dM_{\beta_0}(s)+o_P(1)$, and thus $\hat{\Lambda}$ is asymptotically linear with influence function

$$\begin{aligned}\tilde{\psi}(t) &= \int_0^t \frac{dM_{\beta_0}(s)}{PY(s)e^{\beta_0' Z}} \\ &\quad -\left\{\int_0^\tau [Z-E(s,\beta_0)]'\,dM_{\beta_0}(s)\right\}V^{-1}(\beta_0)\int_0^t E(s,\beta_0)d\Lambda_0(s).\end{aligned} \tag{4.7}$$

Since $\{\tilde{\psi}(t): t\in T\}$ is Donsker, $\sqrt{n}(\hat{\Lambda}-\Lambda)$ converges weakly in $D[0,\tau]$ to a tight, mean zero Gaussian process $\mathcal{Z}$ with covariance $P\left[\tilde{\psi}(s)\tilde{\psi}(t)\right]$. Let $\hat{\psi}(t)$ be $\tilde{\psi}(t)$ with $\hat{\beta}$ and $\hat{\Lambda}$ substituted for $\beta_0$ and $\Lambda_0$, respectively.

After some additional analysis, it can be verified that

$$\sup_{t\in[0,\tau]}\mathbb{P}_n\left[\hat{\psi}(t)-\tilde{\psi}(t)\right]^2 \xrightarrow{\mathrm{P}} 0.$$

Since $\hat{\psi}$ has envelope $kN(\tau)$, for some fixed $k<\infty$, the class $\{\hat{\psi}(s)\hat{\psi}(t): s,t\in[0,\tau]\}$ is Glivenko-Cantelli, and thus $\mathbb{P}_n\left[\hat{\psi}(s)\hat{\psi}(t)\right]$ is uniformly consistent (in probability) for $P\left[\tilde{\psi}(s)\tilde{\psi}(t)\right]$ by the discussion in the beginning of Section 2.2.3 and the smoothness of the involved functions in terms of $\beta$ and $\Lambda$. However, this is not particularly helpful for inference, and the following approach is better. Let $\xi$ be standard normal and independent of the data $X=(N,Y,Z)$, and consider the wild bootstrap $\tilde{\Delta}(t)=\mathbb{P}_n\xi\hat{\psi}(t)$. Although some work is needed, it can be shown that $\sqrt{n}\tilde{\Delta}\overset{\mathrm{P}}{\underset{\xi}{\rightsquigarrow}}\mathcal{Z}$. This is computationally quite simple, since it requires saving $\hat{\psi}_1(t_j),\ldots,\hat{\psi}_n(t_j)$ only at all of the observed jump points $t_1,\ldots,t_{m_n}$, drawing a sample of standard normals $\xi_1,\ldots,\xi_n$, evaluating $\sup_{1\le j\le m_n}\left|\tilde{\Delta}(t)\right|$, and repeating often enough to obtain a reliable estimate of the $(1-\alpha)$-level quantile $\hat{c}_\alpha$. The confidence band $\{\hat{\Lambda}(t)\pm\hat{c}_\alpha: t\in[0,\tau]\}$ thus has approximate coverage $1-\alpha$. A number of modifications of this are possible, including a modification where the width of the band at $t$ is roughly proportional to the variance of $\sqrt{n}(\hat{\Lambda}(t)-\Lambda_0(t))$.

### *4.2.2 The Cox Model*

For the Cox regression model applied to right-censored failure time data, we observe $X = (W, \delta, Z)$, where $W = T \wedge C$, $\delta = 1\{W = T\}$, $Z \in \mathbb{R}^k$ is a regression covariate, $T$ is a right-censored failure time with integrated hazard $e^{\beta' Z}\Lambda(t)$ given the covariate, and where $C$ is a censoring time independent of $T$ given $Z$. We also assume that censoring is uninformative of $\beta$ or $\Lambda$. This is a special case of the general counting process regression model of the previous section, with $N(t) = \delta 1\{W \le t\}$ and $Y(t) = 1\{W \ge t\}$. The consistency of $\hat{\beta}$, a zero of $U_n(\tau, \beta)$, and $\hat{\Lambda}$ both follow from the previous general results, as does the asymptotic normality and validity of the multiplier bootstrap based on the estimated influence function. There are, however, some interesting special features of the Cox model that are of interest, including the martingale structure, the limiting covariance, and the efficiency of the estimators.

First, it is not difficult to show that $M_{\beta_0}(t)$ and $U_n(t, \beta_0)$ are both continuous-time martingales (Fleming and Harrington, 1991; Andersen, Borgan, Gill and Keiding, 1993). This implies that

$$P\left\{\int_0^\tau [Z - E(s, \beta_0)]\, dM_{\beta_0}(s)\right\}^{\otimes 2} = V(\beta_0),$$

and thus the asymptotic limiting variance of $\sqrt{n}(\hat{\beta} - \beta_0)$ is simply $V^{-1}(\beta_0)$. Thus $\hat{\beta}$ is efficient. In addition, the results of this section verify Condition (3.7) and the non-displayed conditions of Theorem 3.1 for $\hat{\ell}_{\tilde{\beta},n}$ defined in (3.8). Verification of (3.6) is left as an exercise (see Exercise 4.6.4). This provides another proof of the efficiency of $\hat{\beta}$. What remains to be verified is that $\hat{\Lambda}$ is also efficient (in the uniform sense). The influence function for $\hat{\Lambda}$ is $\tilde{\psi}$ given in (4.7). We will now use the methods of Section 3.2 to verify that $\tilde{\psi}(u)$ is the efficient influence function for the parameter $\Lambda(u)$, for each $u \in [0, \tau]$. This will imply that $\hat{\Lambda}$ is uniformly efficient for $\Lambda$, since $\sqrt{n}(\hat{\Lambda} - \Lambda)$ converges weakly to a tight, mean zero Gaussian process. This last argument will be vindicated in Theorem 18.9 on Page 341.

The tangent space for the Cox model (for both parameters together) is $\{A(a,b) = \int_0^\tau [Z'a + b(s)]\, dM_\beta(s) : (a,b) \in H\}$, where $H = \mathbb{R}^k \times L_2(\Lambda)$. $A : H \mapsto L_2^0(P_{\beta,\Lambda})$ is thus a score operator for the full model. The natural inner product for pairs of elements in $H$ is $\langle (a,b),(c,d)\rangle_H = a'b + \int_0^\tau c(s)d(s)d\Lambda(s)$, for $(a,b),(c,d) \in H$; and the natural inner product for $g, h \in L_2^0(P_{\beta,\Lambda})$ is $\langle g, h\rangle_{P_{\beta,\Lambda}} = P_{\beta,\Lambda}[gh]$. Hence the adjoint of $A$, $A^*$, satisfies $\langle A(a,b), g\rangle_{P_{\beta,\Lambda}} = \langle (a,b), A^* g\rangle_H$ for all $(a,b) \in H$ and all $g \in L_2^0(P_{\beta,\Lambda})$. It can be shown that $A^* g = \left(P_{\beta,\Lambda}\left[\int_0^\tau Z dM_\beta(s) g\right], P_{\beta,\Lambda}[dM_\beta(t) g]/d\Lambda_0(t)\right)$ satisfies these equations and is thus the adjoint of $A$.

Let the one-dimensional submodel $\{P_t : t \in N_\epsilon\}$ be a perturbation in the direction $A(a,b)$, for some $(a,b) \in H$. In Section 3.2, we showed that $\Lambda_t(u) = \int_0^u (1 + tb(s))d\Lambda(s) + o(t)$, and thus, for the parameter $\chi(P_{\beta,\Lambda}) =$

$\Lambda(u)$, $\partial\chi(P_t)/(\partial t)|_{t=0} = \int_0^u b(s)d\Lambda(s)$. Thus $\langle\tilde{\chi},(a,b)\rangle_H = \int_0^u b(s)d\Lambda(s)$, and therefore $\tilde{\chi} = (0, 1\{s \le v\}) \in H$ ($s$ is the function argument here). Our proof will be complete if we can show that $A^*\tilde{\psi} = \tilde{\chi}$, since this would imply that $\tilde{\psi}(u)$ is the efficient influence function for $\Lambda(u)$. First,

$$\begin{aligned} P\left[\int_0^\tau Z dM_\beta(s)\tilde{\psi}(u)\right] &= P_{\beta,\Lambda}\left\{\int_0^\tau Z dM_\beta(s)\int_0^u \frac{dM_\beta(s)}{P_{\beta,\Lambda}\left[Y(s)e^{\beta' Z}\right]}\right\} \\ &\quad -P_{\beta,\Lambda}\left\{\int_0^\tau Z dM_\beta(s)\int_0^\tau \left[Z - E(s,\beta)\right]' dM_\beta(s)\right\} \\ &\quad \times V^{-1}(\beta)\int_0^u E(s,\beta)d\Lambda(s) \\ &= \int_0^u E(s,\beta)d\Lambda(s) - V(\beta)V^{-1}(\beta)\int_0^u E(s,\beta)d\Lambda(s) \\ &= 0. \end{aligned}$$

Second, it is not difficult to verify that $P_{\beta,\Lambda}\left[dM_\beta(s)\tilde{\psi}(u)\right] = 1\{s \le u\}d\Lambda(s)$, and thus we obtain the desired result. Therefore, both parameter estimators $\hat{\beta}$ and $\hat{\Lambda}$ are uniformly efficient for estimating $\beta$ and $\Lambda$ in the Cox model, since asymptotic tightness plus pointwise efficiency implies uniform efficiency (see Theorem 18.9 in Part III).

## 4.3 The Kaplan-Meier Estimator

In this section, we consider the same right-censored failure time data considered in Section 4.2.2, except that there is no regression covariate. The precise set-up is described in Section 2.2.5. Thus $T$ and $C$ are assumed to be independent, where $T$ has distribution function $F$ and $C$ has distribution function $G$. Assume also that $F(0) = 0$. We denote $P_F$ to be the probability measure for the observed data $X = (W, \delta)$, and allow both $F$ and $G$ to have jumps. Define $S = 1 - F$, $L = 1 - G$, and $\pi(t) = P_F Y(t)$, and let $\tau \in (0, \infty)$ satisfy $F(\tau) > 0$ and $\pi(\tau) > 0$. Also define $\hat{\Lambda}(t) = \int_0^t \left[\mathbb{P}_n Y(s)\right]^{-1} \mathbb{P}_n dN(s)$. The Kaplan-Meier estimator $\hat{S}$ has the product integral form $\hat{S}(t) = \prod_{0<s\le t}\left[1 - d\hat{\Lambda}(s)\right]$.

We have already established in Section 2.2.5, using the self-consistency representation of $\hat{S}$, that $\hat{S}$ is uniformly consistent for $S$ over $[0, \tau]$ and that $\sqrt{n}\left[\hat{S} - S\right]$ converges weakly in $D[0, \tau]$ to a tight, mean zero Gaussian process. In this section, we verify that $\hat{S}$ is also uniformly efficient. We first derive the influence function $\tilde{\psi}$ for $\hat{S}$, and then show that this satisfies the appropriate version of the adjoint formula (3.5) for estimating $S(u)$, for each $u \in [0, \tau]$. This pointwise efficiency will then imply uniform efficiency because of the weak convergence of $\sqrt{n}\left[\hat{S} - S\right]$.

Standard calculations (Fleming and Harrington, 1991, Chapter 3) reveal that

$$\begin{aligned}\hat{S}(u) - S(u) &= -\hat{S}(u)\int_0^u \frac{S(v-)}{\hat{S}(v)}\left\{d\hat{\Lambda}(v) - d\Lambda(v)\right\} \\ &= -\hat{S}(u)\int_0^u 1\{\mathbb{P}_n Y(v) > 0\}\frac{S(v-)}{\hat{S}(v)}\left\{\frac{\mathbb{P}_n dM(v)}{\mathbb{P}_n Y(v)}\right\} \\ &\quad +\hat{S}(u)\int_0^u 1\{\mathbb{P}_n Y(v) = 0\}\frac{S(v-)}{\hat{S}(v)}d\Lambda(v) \\ &= A_n(u) + B_n(u),\end{aligned}$$

where $M(v) = N(v) - \int_0^v Y(s)d\Lambda(s)$ is a martingale. Since $\pi(\tau) > 0$, $B_n(u) = o_P(n^{-1/2})$. Using martingale methods, it can be shown that $A_n(u) = \mathbb{P}_n\check{\psi}(u) + o_P(n^{1/2})$, where

$$\check{\psi}(u) = -S(u)\int_0^u \frac{1}{1-\Delta\Lambda(v)}\left\{\frac{dM(v)}{\pi(v)}\right\}.$$

Since all error terms are uniform for $u \in [0,\tau]$, we obtain that $\hat{S}$ is asymptotically linear, with influence function $\check{\psi}$. As an element of $L_2^0(P_F)$, $\check{\psi}(u)(W,\delta) = \check{g}(W,\delta)$, where

$$\begin{aligned}\check{g}(W,1) &= -S(u)\left[\frac{1\{W \le u\}}{[1-\Delta\Lambda(W)]\,\pi(W)} - \int_0^u \frac{1\{W \ge s\}d\Lambda(s)}{[1-\Delta\Lambda(s)]\,\pi(s)}\right] \text{ and} \\ \check{g}(W,0) &= -\int_0^u \frac{1\{W \ge s\}d\Lambda(s)}{[1-\Delta\Lambda(s)]\,\pi(s)}.\end{aligned}$$

For each $h \in L_2^0(F)$, there exists a one-dimensional submodel $\{F_t : t \in N_\epsilon\}$, with $F_t(v) = \int_0^v (1 + th(s))dF(s) + o(t)$. This is clearly the maximal tangent set for $F$. This collection of submodels can be shown to generate the tangent set for the observed data model $\{P_F : F \in \mathcal{D}\}$, where $\mathcal{D}$ is the collection of all failure time distribution functions with $F(0) = 0$, via the score operator $A : L_2^0(F) \mapsto L_2^0(P_F)$ defined by $(Ah)(W,\delta) = \delta h(W) + (1-\delta)\int_W^\infty h(v)dF(v)/S(W)$. For $a, b \in L_2^0(F)$, let $\langle a,b\rangle_F = \int_0^\infty a(s)b(s)dF(s)$; and for $h_1, h_2 \in L_2^0(P_F)$, let $\langle h_1,h_2\rangle_{P_F} = P_F[h_1h_2]$. Thus the adjoint of $A$ must satisfy $\langle Ah, g\rangle_{P_F} = \langle h, A^*g\rangle_F$. Accordingly,

$$\begin{aligned}P_F\left[(Ah)g\right] &= P_F\left[\delta h(W)g(W,1) + (1-\delta)\frac{\int_W^\infty h(v)dF(v)}{S(W)}g(W,0)\right] \\ &= \int_0^\infty g(v,1)L(v-)h(v)dF(v) + \int_0^\infty \frac{\int_w^\infty h(v)dF(v)}{S(w)}S(w)dG(w) \\ &= \int_0^\infty g(v,1)L(v-)h(v)dF(v) - \int_0^\infty \int_{[v,\infty]} g(w,0)dG(w)h(v)dF(v),\end{aligned}$$

by the fact that $\int_s^\infty h(v)dF(v) = -\int_0^s h(v)dF(v)$ and by changing the order of integration on the right-hand-side. Thus $A^*g(v) = g(v,1)L(v-) - \int_{[v,\infty]} g(s,0)dG(s)$.

With $S_t(u) = 1 - F_t(u)$ based on a submodel perturbed in the direction $h \in L_2^0(F)$, we have that $\partial S_t(u)/(\partial t)|_{t=0} = \int_u^\infty h(v)dF(v) = -\int_0^u h(v)dF(v) = \langle \tilde{\chi}, h \rangle_F$, where $\tilde{\chi} = -1\{v \le u\}$. We now verify that $A^*\left[\check{\psi}(u)\right] = \tilde{\chi}$, and thereby prove that $\hat{S}(u)$ is efficient for estimating the parameter $S(u)$, since it can be shown that $\check{\psi} \in R(A)$. We now have

(4.8)

$$\begin{aligned}\left(A^*\left[\check{\psi}(u)\right]\right)(v) &= \check{\psi}(u)(v,1)L(v-) - \int_{[v,\infty]} \check{\psi}(u)(s,0)dG(s) \\ &= -\frac{S(u)}{S(v)}1\{v \le u\} + S(u)L(v-)\int_0^u \frac{1\{v \ge s\}d\Lambda(s)}{[1-\Delta\Lambda(s)]\,\pi(s)} \\ &\quad - \int_{[v,\infty]} \left\{ S(u)\int_0^u \frac{1\{s \ge r\}d\Lambda(r)}{[1-\Delta\Lambda(r)]\,\pi(r)} \right\} dG(s).\end{aligned}$$

Since $\int_{[v,\infty]} 1\{s \ge r\}dG(s) = L([v \vee r]-) = 1\{v \ge r\}L(v-) + 1\{v < r\}L(r-)$, we now have that

$$\begin{aligned}(4.8) &= -\frac{S(u)}{S(v)}1\{v \le u\} - S(u)\int_0^u \frac{1\{v < r\}L(r-)d\Lambda(r)}{[1-\Delta\Lambda(r)]\,\pi(r)} \\ &= -1\{v \le u\}\left[\frac{1}{S(v)} + \int_v^u \frac{dF(r)}{S(r)S(r-)}\right] S(u) \\ &= -1\{v \le u\}.\end{aligned}$$

Thus (3.5) is satisfied, and we obtain the result that $\hat{S}$ is pointwise and, therefore, uniformly efficient for estimating $S$.

## 4.4 Efficient Estimating Equations for Regression

We now consider a generalization of the conditionally mean zero residual linear regression model considered previously. A typical observation is assumed to be $X = (Y, Z)$, where $Y = g_\theta(Z) + e$, $\mathrm{E}\{e|Z\} = 0$, $Z, \theta \in \mathbb{R}^k$, and $g_\theta(Z)$ is a known, sufficiently smooth function of $\theta$. In addition to linear regression, generalized linear models—as well as many nonlinear regression models—fall into this structure. We assume that $(Z, e)$ has a density $\eta$, and, therefore, that the observation $(Y, Z)$ has density $\eta(y - g_\theta(z), z)$ with the only restriction being that $\int_{\mathbb{R}} e\eta(e,z)de = 0$. As we observed in Section 3.2, these conditions force the score functions for $\eta$ to be all square-integrable functions $a(e,z)$ which satisfy

$$\mathrm{E}\{ea(e,Z)|Z\} = \frac{\int_{\mathbb{R}} ea(e,Z)\eta(e,Z)de}{\int_{\mathbb{R}} \eta(e,Z)de} = 0$$

almost surely. As also demonstrated in Section 3.2, the above equality implies that the tangent space for $\eta$ is the orthocomplement in $L_2^0(P_{\theta,\eta})$ of the set $\mathcal{H}$ of all functions of the form $eh(Z)$, where $\mathrm{E}h^2(Z) < \infty$.

Hence, as pointed out previously, the efficient score for $\theta$ is obtained by projecting the ordinary score $\dot{\ell}_{\theta,\eta}(e,z) = -[\dot{\eta}_1(e,z)/\eta(e,z)]\,\dot{g}_\theta(z)$ onto $\mathcal{H}$, where $\dot{\eta}_1$ is the derivative with respect to the first argument of $\eta$ and $\dot{g}_\theta$ is the derivative of $g_\theta$ with respect to $\theta$. Of course, we are assuming that these derivatives exist and are square-integrable. Since the projection of an arbitrary $b(e,z)$ onto $\mathcal{H}$ is $e\mathrm{E}\{eb(e,Z)|Z\}/V(Z)$, where $V(z) \equiv \mathrm{E}\{e^2|Z = z\}$, the efficient score for $\theta$ is

$$\tilde{\ell}_{\theta,\eta}(Y,Z) = -\frac{\dot{g}_\theta(Z)e\int_{\mathbb{R}}\dot{\eta}_1(u,Z)udu}{V(Z)\int_{\mathbb{R}}\eta u,Zdu} = \frac{\dot{g}_\theta(Z)(Y-g_\theta(Z))}{V(Z)}. \tag{4.9}$$

This, of course, implies the result in Section 3.2 for the special case of linear regression.

In practice, the form of $V(Z)$ is typically not known and needs to be estimated. Let this estimator be denoted $\hat{V}(Z)$. It can be shown that, even if $\hat{V}$ is not consistent but converges to $\tilde{V} \neq V$, the estimating equation

$$\hat{S}_n(\theta) \equiv \mathbb{P}_n\left[\frac{\dot{g}_\theta(Z)(Y-g_\theta(Z))}{\hat{V}(Z)}\right]$$

is approximately equivalent to the estimating equation

$$\tilde{S}_n(\theta) \equiv \mathbb{P}_n\left[\frac{\dot{g}_\theta(Z)(Y-g_\theta(Z))}{\tilde{V}(Z)}\right],$$

both of which can still yield $\sqrt{n}$ consistent estimators of $\theta$. The closer $\tilde{V}$ is to $V$, the more efficient will be the estimator based on solving $\hat{S}_n(\theta) = 0$. Another variant of the question of optimality is "for what choice of $\tilde{V}$ will the estimator obtained by solving $\tilde{S}_n(\theta) = 0$ yield the smallest possible variance?" Godambe (1960) showed that, for univariate $\theta$, the answer is $\tilde{V} = V$. This result is not based on semiparametric efficiency analysis, but is obtained from minimizing the limiting variance of the estimator over all "reasonable" choices of $\tilde{V}$.

For any real, measurable function $w$ of $z \in \mathbb{R}^k$ (measurable on the probability space for $Z$), let

$$S_{n,w}(\theta) \equiv \mathbb{P}_n\left[\dot{g}_\theta(Z)(Y-g_\theta(Z))w(Z)/V(Z)\right],$$

$U(z) \equiv \dot{g}(z)\dot{g}'(z)/V(z)$, and let $\hat{\theta}_{n,w}$ be a solution of $S_{n,w}(\theta) = 0$. Standard methods can be used to show that if both $\mathrm{E}[U(Z)]$ and $\mathrm{E}[U(Z)w(Z)]$ are positive definite, then the limiting variance of $\sqrt{n}(\hat{\theta}_{n,w} - \theta)$ is

$$\{\mathrm{E}[U(Z)w(Z)]\}^{-1}\left\{\mathrm{E}[U(Z)w^2(Z)]\right\}\{\mathrm{E}[U(Z)w(Z)]\}^{-1}.$$

The following proposition yields Godambe's (1960) result generalized for arbitrary $k \geq 1$:

PROPOSITION 4.3 *Assume* $\mathrm{E}[U(Z)]$ *is positive definite. Then, for any real,* $Z$*-measurable function* $w$ *for which* $\mathrm{E}[U(Z)w(Z)]$ *is positive definite,*

$$C_{0,w} \equiv \{\mathrm{E}[U(Z)w(Z)]\}^{-1} \mathrm{E}[U(Z)w^2(Z)] \{\mathrm{E}[U(Z)w(Z)]\}^{-1} - \{\mathrm{E}[U(Z)]\}^{-1}$$

*is positive semidefinite.*

**Proof.** Define $B(z) \equiv \{\mathrm{E}[U(Z)w(Z)]\}^{-1} w(z) - \{\mathrm{E}[U(Z)]\}^{-1}$, and note that $C_w(Z) \equiv B(Z)\mathrm{E}[U(Z)]B'(Z)$ must therefore be positive semidefinite. The desired result now follows since $\mathrm{E}[C_w(Z)] = C_{0,w}$.□

Note that when $A - B$ is positive semidefinite, for two $k \times k$ variance matrices $A$ and $B$, we know that $B$ is the smaller variance. This follows since, for any $v \in \mathbb{R}^k$, $v'Av \geq v'Bv$. Thus the choice $w = 1$ will yield the minimum variance, or, in other words, the choice $\tilde{V} = V$ will yield the lowest possible variance for estimators asymptotically equivalent to the solution of $\tilde{S}_n(\theta) = 0$.

We now verify that estimation based on solving $\hat{S}_n(\theta) = 0$ is asymptotically equivalent to estimation based on solving $\tilde{S}_n(\theta) = 0$, under reasonable regularity conditions. Assume that $\hat{\theta}$ satisfies $\hat{S}_n(\hat{\theta}) = o_P(n^{-1/2})$ and that $\hat{\theta} \xrightarrow{\mathrm{P}} \theta$. Assume also that for every $\epsilon > 0$, there exists a $P$-Donsker class $\mathcal{G}$ such that the inner probability that $\hat{V} \in \mathcal{G}$ is $> 1 - \epsilon$ for all $n$ large enough and all $\epsilon > 0$, and that for some $\tau > 0$, the class

$$\mathcal{F}_1 = \left\{ \frac{\dot{g}_{\theta_1}(Z)(Y - g_{\theta_2}(Z))}{W(Z)} : \|\theta_1 - \theta\| \leq \tau, \|\theta_2 - \theta\| \leq \tau, W \in \mathcal{G} \right\}$$

is $P$-Donsker, and the class

$$\mathcal{F}_2 = \left\{ \frac{\dot{g}_{\theta_1}(Z)\dot{g}'_{\theta_2}(Z)}{W(Z)} : \|\theta_1 - \theta\| \leq \tau, \|\theta_2 - \theta\| \leq \tau, W \in \mathcal{G} \right\}$$

is $P$-Glivenko-Cantelli. We also need that

$$P \left[ \frac{\dot{g}_{\hat{\theta}}(Z)}{\hat{V}(Z)} - \frac{\dot{g}_{\theta}(Z)}{\tilde{V}(Z)} \right]^2 \xrightarrow{\mathrm{P}} 0, \tag{4.10}$$

and, for any $\tilde{\theta} \xrightarrow{\mathrm{P}} \theta$,

$$\left| P \frac{\dot{g}_{\hat{\theta}}(Z)\dot{g}'_{\tilde{\theta}}(Z)}{\hat{V}(Z)} - P \frac{\dot{g}_{\theta}(Z)\dot{g}'_{\theta}(Z)}{\tilde{V}(Z)} \right| \xrightarrow{\mathrm{P}} 0. \tag{4.11}$$

We have the following lemma:

LEMMA 4.4 *Assume that* $\mathrm{E}[Y|Z=z]=g_\theta(z)$, *that*

$$U_0 \equiv \mathrm{E}\left[\frac{\dot{g}_\theta(Z)\dot{g}'_\theta(Z)}{V(Z)}\right]$$

*is positive definite, that* $\hat{\theta}$ *satisfies* $\hat{S}_n(\hat{\theta})=o_P(n^{-1/2})$, *and that* $\hat{\theta}\xrightarrow{\mathrm{P}}\theta$. *Suppose also that* $\mathcal{F}_1$ *is* $P$*-Donsker, that* $\mathcal{F}_2$ *is* $P$*-Glivenko-Cantelli, and that both (4.10) and (4.11) hold for any* $\tilde{\theta}\xrightarrow{\mathrm{P}}\theta$. *Then* $\sqrt{n}(\hat{\theta}-\theta)$ *is asymptotically normal with variance*

$$U_0^{-1}\mathrm{E}\left[\frac{\dot{g}_\theta(Z)\dot{g}'_\theta(Z)V(Z)}{\tilde{V}^2(Z)}\right]U_0^{-1}.$$

*Moreover, if* $\tilde{V}=V$, $Z$*-almost surely, then* $\hat{\theta}$ *is optimal in the sense of Proposition 4.3.*

**Proof.** For any $h\in\mathbb{R}^k$,

$$\begin{aligned} o_P(n^{-1/2}) &= h'\mathbb{P}_n\left[\frac{\dot{g}_{\hat{\theta}}(Z)(Y-g_{\hat{\theta}}(Z))}{\hat{V}(Z)}\right] = h'\mathbb{P}_n\left[\frac{\dot{g}_\theta(Z)(Y-g_\theta(Z))}{\tilde{V}(Z)}\right] \\ &\quad + h'\mathbb{P}_n\left[\left\{\frac{\dot{g}_{\hat{\theta}}(Z)}{\hat{V}(Z)}-\frac{\dot{g}_\theta(Z)}{\tilde{V}(Z)}\right\}(Y-g_\theta(Z))\right] \\ &\quad - h'\mathbb{P}_n\left[\frac{\dot{g}_{\hat{\theta}}(Z)\dot{g}'_{\tilde{\theta}}(Z)}{\hat{V}(Z)}(\hat{\theta}-\theta)\right], \end{aligned}$$

where $\tilde{\theta}$ is on the line segment between $\theta$ and $\hat{\theta}$. Now the conditions of the theorem can be seen to imply that

$$\sqrt{n}(\hat{\theta}-\theta)=\sqrt{n}U_0^{-1}\mathbb{P}_n\left\{\frac{\dot{g}_\theta(Z)(Y-g_\theta(Z))}{\tilde{V}(Z)}\right\}+o_P(1),$$

and the first conclusion of the lemma follows. The final conclusion now follows directly from Proposition 4.3.□

The conditions of Lemma 4.4 are easily satisfied for a variety of regression models and variance estimators $\hat{V}$. For example, if $g_\theta(Z)=(1+e^{-\theta' Z})^{-1}$ is the conditional expectation of a Bernoulli outcome $Y$, given $Z$, and both $\theta$ and $Z$ are assumed to be bounded, then all the conditions are easily satisfied with $\hat{\theta}$ being a zero of $\hat{S}_n$ and $\hat{V}(Z)=g_{\hat{\theta}}(Z)\left[1-g_{\hat{\theta}}(Z)\right]$, provided $\mathrm{E}[ZZ']$ is positive definite. Note that for this example, $\hat{\theta}$ is also semiparametric efficient. We now consider two additional examples in some detail. The first example is a special case of the semiparametric model discussed at the beginning of this section. The model is simple linear regression but with an unspecified form for $V(Z)$. Note that for general $g_\theta$, if the conditions of Lemma 4.4 are satisfied for $\tilde{V}=V$, then the estimator $\hat{\theta}$ is semiparametric

efficient by the form of the efficient score given in (4.9). The second example considers estimation of a semiparametric Poisson mixture regression model, where the mixture induces extra-Poisson variation. We will develop an optimal estimating equation procedure in the sense of Proposition 4.3. Unfortunately, it is unclear in this instance how to strengthen this result to obtain semiparametric efficiency.

### *4.4.1 Simple Linear Regression*

Consider simple linear regression based on a univariate $Z$. Let $\theta = (\alpha, \beta)' \in \mathbb{R}^2$ and assume $g_\theta(Z) = \alpha + \beta Z$. We also assume that the support of $Z$ is a known compact interval $[a, b]$. We will use a modified kernel method of estimating $V(z) \equiv E[e^2|Z = z]$. Let the kernel $L : \mathbb{R} \mapsto [0, 1]$ satisfy $L(x) = 0$ for all $|x| > 1$ and $L(x) = 1 - |x|$ otherwise, and let $h \leq (b-a)/2$ be the bandwidth for this kernel. For the sample $(Y_1, Z_1), \ldots, (Y_n, Z_n)$, let $\hat{\theta} = (\hat{\alpha}, \hat{\beta})'$ be the usual least-squares estimator of $\theta$, and let $\hat{e}_i = Y_i - \hat{\alpha} - \hat{\beta} Z_i$ be the estimated residuals. Let $\hat{F}_n(u)$ be the empirical distribution of $Z_1, \ldots, Z_n$, and let $\hat{H}_n(u) \equiv n^{-1} \sum_{i=1}^n \hat{e}_i^2 1\{Z_i \leq u\}$. We will denote $F$ as the true distribution of $Z$, with density $f$, and also define $H(z) \equiv \int_a^z V(u) dF(u)$. Now define

$$\hat{V}(z) \equiv \frac{\int_{\mathbb{R}} h^{-1} L\left(\frac{z-u}{h}\right) \hat{H}_n(du)}{\int_{\mathbb{R}} h^{-1} L\left(\frac{z-u}{h}\right) \hat{F}_n(du)},$$

for $z \in [a + h, b - h]$, and let $\hat{V}(z) = \hat{V}(a + h)$ for all $z \in [a, a + h)$ and $\hat{V}(z) = \hat{V}(b - h)$ for all $z \in (b - h, b]$.

We need to assume in addition that both $F$ and $V$ are twice differentiable with second derivatives uniformly bounded on $[a, b]$, that for some $M < \infty$ we have $M^{-1} \leq f(z), V(z) \leq M$ for all $a \leq x \leq b$, and that the possibly data-dependent bandwidth satisfies $h = o_P(1)$ and $h^{-1} = o_P(n^{1/4})$. If we let $U_i \equiv (1, Z_i)'$, $i = 1, \ldots, n$, then $\hat{H}_n(z) =$

$$\begin{aligned}
n^{-1} \sum_{i=1}^n \hat{e}_i^2 1\{Z_i \leq z\} &= n^{-1} \sum_{i=1}^n \left[e_i - (\hat{\theta} - \theta)' U_i\right]^2 1\{Z_i \leq z\} \\
&= n^{-1} \sum_{i=1}^n e_i^2 1\{Z_i \leq z\} \\
&\quad -2(\hat{\theta} - \theta)' n^{-1} \sum_{i=1}^n U_i e_i 1\{Z_i \leq z\} \\
&\quad +(\hat{\theta} - \theta)' n^{-1} \sum_{i=1}^n U_i U_i' 1\{Z_i \leq z\} (\hat{\theta} - \theta) \\
&= A_n(z) - B_n(z) + C_n(z).
\end{aligned}$$

In Chapter 9, we will show in an exercise (see Exercise 9.6.8) that $\mathcal{G}_1 \equiv \{Ue \cdot 1\{Z \leq z\}, z \in [a,b]\}$ is Donsker. Hence $\|\mathbb{P}_n - P\|_{\mathcal{G}_1} = O_P(n^{-1/2})$. Since also $\mathrm{E}[e|Z] = 0$ and $\|\hat{\theta} - \theta\| = O_P(n^{-1/2})$, we now have that

$$\sup_{z \in [a,b]} |B_n(z)| = O_P(n^{-1}).$$

By noting that $\|U\|$ is bounded under our assumptions, we also obtain that $\sup_{z\in[a,b]} |C_n(z)| = O_P(n^{-1})$. In Exercise 9.6.8 in Chapter 9, we will also verify that $\mathcal{G}_2 \equiv \{e^2 \cdot 1\{Z \leq z\}, z \in [a,b]\}$ is Donsker. Hence $\sup_{z\in[a,b]} |A_n(z) - H(z)| = O_P(n^{-1/2})$, and thus also $\sup_{z\in[a,b]} |\hat{H}_n(z) - H(z)| = O_P(n^{-1/2})$. Standard results also verify that $\sup_{z\in[a,b]} |\hat{F}_n(z) - F(z)| = O_P(n^{-1/2})$.

Let $D_n \equiv \hat{H}_n - H$, $\dot{L}$ be the derivative of $L$, and note that for any $z \in [a+h, b-h]$ we have by integration by parts and by the form of $\dot{L}$ that

$$\int_{\mathbb{R}} h^{-1} L\left(\frac{z-u}{h}\right) D_n(du) = -\int_{z-h}^{z+h} D_n(u) h^{-2} \dot{L}\left(\frac{z-u}{h}\right) du.$$

Thus

$$\left| \int_{\mathbb{R}} h^{-1} L\left(\frac{z-u}{h}\right) D_n(du) \right| \leq h^{-1} \sup_{z\in[a,b]} \left| \hat{H}_n(z) - H(z) \right|.$$

Since the right-hand-side does not depend on $z$, and by the result of the previous paragraph, we obtain that the supremum of the left-hand-side over $z \in [a+h, b-h]$ is $O_P(h^{-1}n^{-1/2})$. Letting $\dot{H}$ be the derivative of $H$, we save it as an exercise (see Exercise 9.6.8 again) to verify that both

$$\sup_{z\in[a+h,b-h]} \left| \int_{\mathbb{R}} h^{-1} L\left(\frac{z-u}{h}\right) H(du) - \dot{H}(z) \right| = O(h)$$

and $\left(\sup_{z\in[a,a+h)} \left|\dot{H}(z) - \dot{H}(a+h)\right|\right) \vee \left(\sup_{z\in(b-h,b]} \left|\dot{H}(z) - \dot{H}(b-h)\right|\right) = O(h)$. Hence

$$\hat{R}_n(z) \equiv \begin{cases} \int_{\mathbb{R}} h^{-1} L\left(\frac{z-u}{h}\right) \hat{H}_n(du) \text{ for } z \in [a+h, b-h], \\ \hat{R}_n(a+h) \text{ for } z \in [a, a+h), \\ \hat{R}_n(b-h) \text{ for } z \in (b-h, b] \end{cases}$$

is uniformly consistent for $\dot{H}$ with uniform error $O_P(h + h^{-1}n^{-1/2}) = o_P(1)$.

Similar, but somewhat simpler arguments compared to those in the previous paragraph, can also be used to verify that

$$\hat{Q}_n(z) \equiv \begin{cases} \int_{\mathbb{R}} h^{-1} L\left(\frac{z-u}{h}\right) \hat{F}_n(du) \text{ for } z \in [a+h, b-h], \\ \hat{Q}_n(a+h) \text{ for } z \in [a, a+h), \\ \hat{Q}_n(b-h) \text{ for } z \in (b-h, b] \end{cases}$$

is uniformly consistent for $f$ also with uniform error $O_P(h + h^{-1}n^{-1/2}) = o_P(1)$. Since $f$ is bounded below, we now have that $\sup_{z\in[a,b]} |\hat{V}(z) - V(z)| = o_P(1)$. Since $\dot{g}_\theta(Z) = (1, Z)'$, we have established that both (4.10) and (4.11) hold for this example since $V$ is also bounded below.

We will now show that there exists a $k_0 < \infty$ such that the probability that "$\hat{V}$ goes below $1/k_0$ or the first derivative of $\hat{V}$ exceeds $k_0$" goes to zero as $n \to \infty$. If we let $\mathcal{G}$ be the class of all functions $q : [a, b] \mapsto [k_0^{-1}, k_0]$ such that the first derivative of $q$, $\dot{q}$, satisfies $|\dot{q}| \leq k_0$, then our result will imply that the inner probability that $\hat{V} \in \mathcal{G}$ is $> 1-\epsilon$ for all $n$ large enough and all $\epsilon > 0$. An additional exercise in Chapter 9 (see again Exercise 9.6.8) will then show that, for this simple linear regression example, the class $\mathcal{F}_1$ defined above is Donsker and $\mathcal{F}_2$ defined above is Glivenko-Cantelli. Thus Lemma 4.4 applies. This means that if we first use least-squares estimators of $\alpha$ and $\beta$ to construct the estimator $\hat{V}$, and then compute the "two-stage" estimator

$$\tilde{\theta} \equiv \left[\sum_{i=1}^{n} \frac{U_i U_i'}{\hat{V}(Z_i)}\right]^{-1} \sum_{i=1}^{n} \frac{U_i Y_i}{\hat{V}(Z_i)},$$

then this $\tilde{\theta}$ will be efficient for $\theta$.

Since $\hat{V}$ is uniformly consistent, the only thing remaining to show is that the derivative of $\hat{V}$, denoted $\dot{V}_n$, is uniformly bounded. Note that the derivative of $\hat{R}_n$, which we will denote $\dot{R}_n$, satisfies the following for all $z \in [a + h, b - h]$:

$$\begin{aligned} \dot{R}_n(z) &= -\int_{\mathbb{R}} h^{-2} \dot{L}\left(\frac{z-u}{h}\right) \hat{H}_n(du) \\ &= h^{-2}\left[\hat{H}_n(z) - \hat{H}_n(z-h) - \hat{H}_n(z+h) + \hat{H}_n(z)\right] \\ &= O_P(h^{-2}n^{-1/2}) + h^{-2}\left[H(z) - H(z-h) - H(z+h) + H(z)\right], \end{aligned}$$

where the last equality follows from the previously established fact that $\sup_{z\in[a+h,b-h]} \left|\hat{H}_n(z) - H(z)\right| = O_P(n^{-1/2})$. Now the uniform boundedness of the second derivative of $H$ ensures that $\sup_{z\in[a+h,b-h]} \left|\dot{H}_n(z)\right| = O_P(1)$. Similar arguments can be used to establish that the derivative of $\hat{Q}_n$, which we will denote $\dot{Q}_n$, satisfies $\sup_{z\in[a+h,b-h]} \left|\dot{Q}_n(z)\right| = O_P(1)$. Now we have uniform boundedness in probability of $\dot{V}_n$ over $[a + h, b - h]$. Since $\hat{V}$ does not change over either $[a, a + h)$ or $(b - h, b]$, we have also established that $\sup_{z\in[a,b]} \left|\dot{V}_n(z)\right| = O_P(1)$, and the desired results follow.

### *4.4.2 A Poisson Mixture Regression Model*

In this section, we consider a Poisson mixture regression model in which the nuisance parameter is not $\sqrt{n}$ consistent in the uniform norm. Given a regression vector $Z \in \mathbb{R}^k$ and a nonnegative random quantity $W \in \mathbb{R}$, the observation $Y$ is Poisson with parameter $We^{\beta' Z}$, for some $\beta \in \mathbb{R}^k$. We only observe the pair $(Y, Z)$. Thus the density of $Y$ given $Z$ is

$$Q_{\beta,G}(y) = \int_0^\infty \frac{e^{-we^{\beta' Z}}[we^{\beta' Z}]^y}{y!} dG(w),$$

where $G$ is the unknown distribution function for $W$. For identifiability, we assume that one of the components of $Z$ is the constant 1 and that the expectation of $W$ is also 1. We denote the joint distribution of the observed data as $P_{\beta,G}$, and assume that $Z$ is bounded, $P_{\beta,G}[ZZ']$ is full rank, and that $P_{\beta,G}Y^2 < \infty$.

In this situation, it is not hard to verify that

$$\begin{aligned} V(z) &= \mathrm{E}\left\{ \left[Y - e^{\beta' Z}\right]^2 \middle| Z = z \right\} \\ &= \mathrm{E}\left[ \mathrm{E}\left\{ \left[Y - Ue^{\beta' Z}\right]^2 \middle| U, Z = z \right\} + (U-1)^2 e^{2\beta' z} \right] \\ &= e^{\beta' z} + \sigma^2 e^{2\beta' z}, \end{aligned}$$

where $\sigma^2 \equiv \mathrm{E}[U-1]^2$. Let $\tilde{\beta}$ be the estimator obtained by solving

$$\mathbb{P}_n \left[ Z(Y - e^{\beta' Z}) \right] = 0.$$

Standard arguments reveal that $\sqrt{n}(\tilde{\beta} - \beta)$ is asymptotically mean zero Gaussian with finite variance matrix. Relatively simple calculations also reveal that $\mathrm{E}[Y(Y-1)|Z=z] = \int_0^\infty u^2 dG(u) e^{2\beta' Z} = (\sigma^2+1)e^{2\beta' Z}$. Hence $\hat{\sigma}^2 \equiv -1 + n^{-1}\sum_{i=1}^n e^{-2\hat{\beta}' Z_i} Y_i(Y_i - 1)$ will satisfy $\hat{\sigma}^2 = \sigma^2 + O_P(n^{-1/2})$. Now let $\hat{\beta}$ be the solution of

$$\mathbb{P}_n \left[ \frac{Z(Y - e^{\beta' Z})}{\hat{V}(Z)} \right] = 0,$$

where $\hat{V}(z) \equiv e^{\tilde{\beta}' z} + \hat{\sigma}^2 e^{2\tilde{\beta}' z}$. It is left as an exercise (see Exercise 4.6.7) to verify that $\hat{\beta}$ satisfies the conditions of Lemma 4.4 for $\tilde{V} = V$, and thus the desired optimality is achieved.

## 4.5 Partly Linear Logistic Regression

For the partly linear logistic regression example given in Chapter 1, the observed data are $n$ independent realizations of the random triplet $(Y, Z, U)$,

where $Z \in \mathbb{R}^p$ and $U \in \mathbb{R}$ are covariates which are not linearly dependent, $Y$ is a dichotomous outcome with conditional expectation $\nu[\beta' Z + \eta(U)]$, $\beta \in \mathbb{R}^p$, $Z$ is restricted to a bounded set, $U \in [0,1]$, $\nu(t) = 1/(1+e^{-t})$, and where $\eta$ is an unknown smooth function. Hereafter, for simplicity, we will also assume that $p = 1$. We further assume, for some integer $k \geq 1$, that the first $k-1$ derivatives of $\eta$ exist and are absolutely continuous with $J^2(\eta) = \int_0^1 \left[\eta^{(k)}(t)\right]^2 dt < \infty$. To estimate $\beta$ and $\eta$ based on an i.i.d. sample $X_i = (Y_i, Z_i, U_i), i = 1, \ldots, n$, we use the following penalized log-likelihood:

$$\tilde{L}_n(\beta, \eta) = n^{-1} \sum_{i=1}^{n} \log p_{\beta,\eta}(X_i) - \hat{\lambda}_n^2 J^2(\eta),$$

where

$$p_{\beta,\eta}(x) = \{\nu[\beta z + \eta(u)]\}^y \{1 - \nu[\beta z + \eta(u)]\}^{1-y}$$

and $\hat{\lambda}_n$ is chosen to satisfy $\hat{\lambda}_n = o_P(n^{-1/4})$ and $\hat{\lambda}_n^{-1} = O_P(n^{k/(2k+1)})$. Denote $\hat{\beta}_n$ and $\hat{\eta}_n$ to be the maximizers of $\tilde{L}_n(\beta, \eta)$, let $P_{\beta,\eta}$ denote expectation under the model, and let $\beta_0$ and $\eta_0$ to be the true values of the parameters.

Consistency of $\hat{\beta}_n$ and $\hat{\eta}_n$ and efficiency of $\hat{\beta}_n$ are established for partly linear generalized linear models in Mammen and van de Geer (1997). We now derive the efficient score for $\beta$ and then sketch a verification that $\hat{\beta}_n$ is asymptotically linear with influence function equal to the efficient influence function. Several technically difficult steps will be reserved for the case studies II chapter, Chapter 15, in Part II of the book. Our approach to proving this diverges only slightly from the approach used by Mamman and van de Geer (1997).

Let $\mathcal{H}$ be the linear space of functions $h : [0,1] \mapsto \mathbb{R}$ with $J(h) < \infty$. For $t \in [0, \epsilon)$ and $\epsilon$ sufficiently small, let $\beta_t = \beta + tv$ and $\eta_t = \eta + th$ for $v \in \mathbb{R}$ and $h \in \mathcal{H}$. If we differentiate the non-penalized log-likelihood, we deduce that the score for $\beta$ and $\eta$, in the direction $(v, h)$, is $(vZ + h(U))(Y - \mu_{\beta,\eta}(Z, U))$, where $\mu_{\beta,\eta}(Z, U) = \nu[\beta Z + \eta(U)]$. Now let

$$h_1(u) = \frac{\mathrm{E}\{Z V_{\beta,\eta}(Z, U) | U = u\}}{\mathrm{E}\{V_{\beta,\eta}(Z, U) | U = u\}},$$

where $V_{\beta,\eta} = \mu_{\beta,\eta}(1 - \mu_{\beta,\eta})$, and assume that $h_1 \in \mathcal{H}$. It can easily be verified that $Z - h_1(U)$ is uncorrelated with any $h(U)$, $h \in \mathcal{H}$, and thus the efficient score for $\beta$ is $\tilde{\ell}_{\beta,\eta}(Y, Z, U) = (Z - h_1(U))(Y - \mu_{\beta,\eta}(Z, U))$. Hence the efficient information for $\beta$ is $\tilde{I}_{\beta,\eta} = P_{\beta,\eta}\left[(Z - h_1(U))^2 V_{\beta,\eta}(Z, U)\right]$ and the efficient influence function is $\tilde{\psi}_{\beta,\eta} = \tilde{I}_{\beta,\eta}^{-1} \tilde{\ell}_{\beta,\eta}$, provided $\tilde{I}_{\beta,\eta} > 0$, which we assume hereafter to be true for $\beta = \beta_0$ and $\eta = \eta_0$.

In order to prove asymptotic linearity of $\hat{\beta}_n$, we need to also assume that $P_{\beta_0,\eta_0}\left[Z - \tilde{h}_1(U)\right]^2 > 0$, where $\tilde{h}_1(u) = \mathrm{E}\{Z | U = u\}$. We prove in

Chapter 15 that $\hat{\beta}_n$ and $\hat{\eta}_n$ are both uniformly consistent for $\beta_0$ and $\eta_0$, respectively, and that

$$\mathbb{P}_n \left[ (\hat{\beta}_n - \beta_0)Z + \hat{\eta}_n(U) - \eta_0(U) \right]^2 = o_P(n^{-1/2}). \tag{4.12}$$

Let $\hat{\beta}_{ns} = \hat{\beta}_n + s$ and $\hat{\eta}_{ns}(u) = \hat{\eta}_n(u) - sh_1(u)$. If we now differentiate $\tilde{L}_n(\hat{\beta}_{ns}, \hat{\eta}_{ns})$ and evaluate at $s = 0$, we obtain

$$\begin{aligned} 0 &= \mathbb{P}_n \left[ (Y - \mu_{\beta_0,\eta_0})(Z - h_1(U)) \right] \\ &\quad - \mathbb{P}_n \left[ (\mu_{\hat{\beta}_n,\hat{\eta}_n} - \mu_{\beta_0,\eta_0})(Z - h_1(U)) \right] - \lambda_n^2 \left\{ \partial J^2(\hat{\eta}_{ns})/(\partial s)|_{s=0} \right\} \\ &= A_n - B_n - C_n, \end{aligned}$$

since $\tilde{L}_n(\beta, \eta)$ is maximized at $\hat{\beta}_n$ and $\hat{\eta}_n$ by definition.

Using (4.12), we obtain

$$\begin{aligned} B_n &= \mathbb{P}_n \left[ V_{\beta_0,\eta_0}(Z,U) \left\{ (\hat{\beta}_n - \beta_0)Z + \hat{\eta}_n(U) - \eta_0(U) \right\} (Z - h_1(U)) \right] \\ &\quad + o_P(n^{-1/2}), \end{aligned}$$

since $\partial \nu(t)/(\partial t) = \nu(1 - \nu)$. By definition of $h_1$, we have

$$P_{\beta_0,\eta_0} \left[ V_{\beta_0,\eta_0}(Z,U)(\hat{\eta}_n(U) - \eta_0(U))(Z - h_1(U)) \right] = 0,$$

and we also have that $P_{\beta_0,\eta_0} \left[ V_{\beta_0,\eta_0}(Z,U)(\hat{\eta}_n(U) - \eta_0(U))(Z - h_1(U)) \right]^2 \xrightarrow{\mathrm{P}} 0$. Thus, if we can establish that for each $\tau > 0$, $\hat{\eta}_n(U) - \eta_0(U)$ lies in a bounded $P_{\beta_0,\eta_0}$-Donsker class with probability $> (1 - \tau)$ for all $n \geq 1$ large enough and all $\tau > 0$, then

$$\mathbb{P}_n \left[ V_{\beta_0,\eta_0}(Z,U)(\hat{\eta}_n(U) - \eta_0(U))(Z - h_1(U)) \right] = o_P(n^{-1/2}),$$

and thus

$$B_n = (\hat{\beta}_n - \beta_0) P_{\beta_0,\eta_0} \left[ V_{\beta_0,\eta_0}(Z,U)(Z - h_1(U))^2 \right] + o_P(n^{-1/2}), \tag{4.13}$$

since products of bounded Donsker classes are Donsker (and therefore also Glivenko-Cantelli), and since

$$P_{\beta_0,\eta_0} \left[ V_{\beta_0,\eta_0}(Z,U) Z (Z - h_1(U)) \right] = P_{\beta_0,\eta_0} \left[ V_{\beta_0,\eta_0}(Z,U)(Z - h_1(U))^2 \right]$$

by definition of $h_1$. Let $\mathcal{H}_c$ be the subset of $\mathcal{H}$ with functions $h$ satisfying $J(h) \leq c$ and $\|h\|_\infty \leq c$. We will show in Chapter 15 that $\{h(U) : h \in \mathcal{H}_c\}$ is indeed Donsker for each $c < \infty$ and also that $J(\hat{\eta}_n) = O_P(1)$. This yields the desired Donsker property, and (4.13) follows.

It is now not difficult to verify that $C_n \leq 2\lambda_n^2 J(\hat{\eta}_n) J(h_1) = o_P(n^{-1/2})$, since $\lambda_n = o_P(n^{-1/4})$ by assumption. Hence $\sqrt{n}(\hat{\beta}_n - \beta_0) = \sqrt{n}\mathbb{P}_n \tilde{\psi}_{\beta_0,\eta_0} + o_P(1)$, and we have verified that $\hat{\beta}_n$ is indeed efficient for $\beta_0$.

## 4.6 Exercises

**4.6.1.** For the linear regression example, verify the inequality (4.3).

**4.6.2.** For the general counting process regression model setting, show that $V(\beta) \geq c\,\mathrm{var}(Z)$, for some $c > 0$, where $V(\beta)$ is given in (4.5).

**4.6.3.** Show how to use Lemma 4.2 to establish (4.6). Hint: Let $A_n(t) = E_n(t, \hat{\beta}) - E(t, \hat{\beta})$ and $B_n(t) = \sqrt{n}\mathbb{P}_n M_{\beta_0}(t)$ for part of it, and let

$$A_n(t) = \mathbb{P}_n \left\{ \int_0^t Y(s) \left[ e^{\hat{\beta}' Z} - e^{\beta_0' Z} \right] d\Lambda_0(s) \right\}$$

and $B_n(t) = \sqrt{n} \left[ E_n(t, \hat{\beta}) - E(t, \hat{\beta}) \right]$ for the other part.

**4.6.4.** For the Cox model example (Section 4.2.2), verify Condition (3.6) of Theorem 3.1 for $\hat{\ell}_{\tilde{\beta},n}$ defined in (3.8).

**4.6.5.** We will see in Section 18.2 (in Theorem 18.8 on Page 339) that, for a regular estimator, it is enough to verify efficiency by showing that the influence function lies in the tangent space for the full model. For the Cox model example, verify that the influence functions for both $\hat{\beta}$ and $\hat{\Lambda}(t)$ (for all $t \in [0, \tau]$) lie in the tangent space for the full model.

**4.6.6.** For the Kaplan-Meier example of Section 4.3, verify that $\check{\psi} \in R(A)$.

**4.6.7.** For the Poisson mixture model example of Section 4.4.2, verify that the conditions of Lemma 4.4 are satisfied for the given choice of $\hat{V}$:

(a) Show that the class $\mathcal{G} \equiv \left\{ e^{t' Z} + s^2 e^{2t' Z} : \|t - \beta\| \leq \epsilon_1, |s^2 - \sigma^2| \leq \epsilon_2 \right\}$ is Donsker for some $\epsilon_1, \epsilon_2 > 0$. Hint: First show that $\{t' Z : \|t - \beta\| \leq \epsilon_1\}$ is Donsker from the fact that the product of two (in this case trivial) bounded Donsker classes is also Donsker. Now complete the proof by using the facts that Lipschitz functions of Donsker classes are Donsker, that products of bounded Donsker classes are Donsker (used earlier), and that sums of Donsker classes are Donsker.

(b) Now complete the verification of Lemma 4.4.

**4.6.8.** For the partly linear logistic regression example of Section 4.5, verify that $(Z - h_1(U))\,(Y - \mu_{\beta_0,\eta_0})$ is uncorrelated with $h(U)(Y - \mu_{\beta_0,\eta_0})$ for all $h \in \mathcal{H}$, where the quantities are as defined in the example.

## 4.7 Notes

Least absolute deviation regression was studied using equicontinuity arguments in Bassett and Koenker (1978). More succinct results based on

empirical processes can be found in Pollard (1991). An excellent discussion of breakdown points and other robustness issues is given in Huber (1981).

# Part II

# Empirical Processes

# 5
# Introduction to Empirical Processes

The goal of Part II is to provide an in depth coverage of the basics of empirical process techniques which are useful in statistics. Chapter 6 presents preliminary mathematical background which provides a foundation for later technical development. The topics covered include metric spaces, outer expectations, linear operators and functional differentiation. The main topics overviewed in Chapter 2 of Part I will then be covered in greater depth, along with several additional topics, in Chapters 7 through 14. Part II finishes in Chapter 15 with several case studies. The main approach is to present the mathematical and statistical ideas in a logical, linear progression, and then to illustrate the application and integration of these ideas in the case study examples. The scaffolding provided by the overview, Part I, should enable the reader to maintain perspective during the sometimes rigorous developments of this section.

Stochastic convergence is studied in Chapter 7. An important aspect of the modes of convergence explored in this book are the notions of outer integrals and outer measure which were mentioned briefly in Section 2.2.1. While many of the standard relationships between the modes of stochastic convergence apply when using outer measure, there are a few important differences that we will examine. While these differences may, in some cases, add complexity to an already difficult asymptotic theory, the gain in breadth of applicability to semiparametric statistical estimators is well worth the trouble. For example, convergence based on outer measure permits the use of the uniform topology for studying convergence of empirical processes with complex index sets. This contrasts with more traditional approaches, which require special topologies that can be harder to use in

applications, such as the Skorohod topology for cadlag processes (see Chapter 3 of Billingsley, 1968).

The main techniques for proving empirical process central limit theorems will be presented in Chapter 8. Establishing Glivenko-Cantelli and Donsker theorems requires bounding expectations involving suprema of stochastic processes. Maximal inequalities and symmetrization techniques are important tools for accomplishing this, and careful measurability arguments are also sometimes needed. Symmetrization involves replacing inequalities for the empirical process $f \mapsto (\mathbb{P}_n - P)f$, $f \in \mathcal{F}$, with inequalities for the "symmetrized" process $n^{-1}\sum_{i=1}^{n} \epsilon_i f(X_i)$, where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher random variables (eg., $\mathrm{P}\{\epsilon_1 = -1\} = \mathrm{P}\{\epsilon_1 = 1\} = 1/2$) independent of $X_1, \ldots, X_n$. Several tools for assessing measurability in statistical applications will also be discussed.

Entropy with bracketing, uniform entropy, and other measures of entropy are essential aspects in all of these results. This is the topic of Chapter 9. The associated entropy calculations can be quite challenging, but the work is often greatly simplified by using Donsker preservation results to build larger Donsker classes from smaller ones. Similar preservation results are available for Glivenko-Cantelli classes.

Bootstrapping of empirical processes, based on multinomial or other Monte Carlo weights, is studied in Chapter 10. The bootstrap is a valuable way to conduct inference for empirical processes because of its broad applicability. In many semiparametric settings, there are no viable alternatives for inference. A central role in establishing validity of the bootstrap is played by multiplier central limit theorems which establish weak convergence of processes of the form $\sqrt{n}\mathbb{P}_n\xi(f(X) - Pf)$, $f \in \mathcal{F}$, where $\xi$ is independent of $X$, has mean zero and variance 1, and $\int_0^\infty \sqrt{P(|\xi| > x)}dx < \infty$.

In Chapter 11, several extensions of empirical process results are presented for function classes which either consist of sequences or change with the sample size $n$, as well as results for independent but not identically distributed data. These results are useful in a number of statistical settings, including asymptotic analysis of the Cramér-von Mises statistic and regression settings where the covariates are assumed fixed or when a biased coin study design (see Wei, 1978, for example) is used. Extensions of the bootstrap for conducting inference in these new situations is also discussed.

Many interesting statistical quantities can be expressed as functionals of empirical processes. The functional delta method, discussed in Chapter 12, can be used to translate weak convergence and bootstrap results for empirical processes to corresponding inference results for these functionals. Most Z- and M- estimators are functionals of empirical processes. For example, under reasonable regularity conditions, the functional which extracts the zero (root) of a Z-estimating equation is sufficiently smooth to permit the delta method to carry over inference results for the estimating equation to the corresponding Z-estimator. The results also apply to M-estimators which can be expressed as approximate Z-estimators.

Z-estimation is discussed in Chapter 13, while M-estimation is discussed in Chapter 14. A key challenge with many important M-estimators is to establish the rate of convergence, especially in settings where the estimators are not $\sqrt{n}$ consistent. This issue was only briefly mentioned in Section 2.2.6 because of the technical complexity of the problem. There are a number of tools which can be used to establish these rates, and several such tools will be studied in Chapter 14. These techniques rely significantly on accurate entropy calculations for the M-estimator empirical process, as indexed by the parameter set, within a small neighborhood of the true parameter.

The case studies presented in Chapter 15 demonstrate that the technical power of empirical process methods facilitates valid inference for flexible models in many interesting and important statistical settings.

# 6
# Preliminaries for Empirical Processes

In this chapter, we cover several mathematical topics that play a central role in the empirical process results we present later. Metric spaces are crucial since they provide the descriptive language by which the most important results about stochastic processes are derived and expressed. Outer expectations, or, more correctly, outer integrals are key to defining and utilizing outer modes of convergence for quantities which are not measurable. Since many statistical quantities of interest are not measurable with respect to the uniform topology, which is often the topology of choice for applications, outer modes of convergence will be the primary approach for stochastic convergence throughout this book. Linear operators and functional derivatives also play a major role in empirical process methods and are key tools for the functional delta method and Z-estimator theory discussed in Chapters 12 and 13.

## 6.1 Metric Spaces

We now introduce a number of concepts and results for metric spaces. Before defining metric spaces, we briefly review topological spaces, $\sigma$-fields, and measure spaces. A collection $\mathcal{O}$ of subsets of a set $X$ is a *topology in* $X$ if:

(i) $\emptyset \in \mathcal{O}$ and $X \in \mathcal{O}$, where $\emptyset$ is the empty set;

(ii) If $U_j \in \mathcal{O}$ for $j = 1, \ldots, m$, then $\bigcap_{j=1,\ldots,m} U_j \in \mathcal{O}$;

(iii) If $\{U_\alpha\}$ is an arbitrary collection of members of $\mathcal{O}$ (finite, countable or uncountable), then $\bigcup_\alpha U_\alpha \in \mathcal{O}$.

When $\mathcal{O}$ is a topology in $X$, then $X$ (or the pair $(X, \mathcal{O})$) is a *topological space*, and the members of $\mathcal{O}$ are called the *open sets* in $X$. For a subset $A \subset X$, the *relative topology* on $A$ consists of the sets $\{A \cap B : B \in \mathcal{O}\}$.

A map $f : X \mapsto Y$ between topological spaces is *continuous* if $f^{-1}(U)$ is open in $X$ whenever $U$ is open in $Y$. A set $B$ in $X$ is *closed* if and only if its complement in $X$, denoted $X - B$, is open. The *closure* of an arbitrary set $E \in X$, denoted $\overline{E}$, is the smallest closed set containing $E$; while the *interior* of an arbitrary set $E \in X$, denoted $E^\circ$, is the largest open set contained in $E$. A subset $A$ of a topological space $X$ is *dense* if $\overline{A} = X$. A topological space $X$ is *separable* if it has a countable dense subset.

A *neighborhood* of a point $x \in X$ is any open set that contains $x$. A topological space is *Hausdorff* if distinct points in $X$ have disjoint neighborhoods. A sequence of points $\{x_n\}$ in a topological space $X$ *converges* to a point $x \in X$ if every neighborhood of $x$ contains all but finitely many of the $x_n$. This convergence is denoted $x_n \to x$. Suppose $x_n \to x$ and $x_n \to y$. Then $x$ and $y$ share all neighborhoods, and $x = y$ when $X$ is Hausdorff. If a map $f : X \mapsto Y$ between topological spaces is continuous, then $f(x_n) \to f(x)$ whenever $x_n \to x$ in $X$. To see this, let $\{x_n\} \subset X$ be a sequence with $x_n \to x \in X$. Then for any open $U \subset Y$ containing $f(x)$, all but finitely many $\{x_n\}$ are in $f^{-1}(U)$, and thus all but finitely many $\{f(x_n)\}$ are in $U$. Since $U$ was arbitrary, we have $f(x_n) \to f(x)$.

We now review the important concept of *compactness*. A subset $K$ of a topological space is *compact* if for every set $A \supset K$, where $A$ is the union of a collection of open sets $\mathcal{S}$, $K$ is also contained in some finite union of sets in $\mathcal{S}$. When the topological space involved is also Hausdorff, then compactness of $K$ is equivalent to the assertion that every sequence in $K$ has a convergent subsequence (converging to a point in $K$). We omit the proof of this equivalence. This result implies that compact subsets of Hausdorff topological spaces are necessarily closed. Note that a compact set is sometimes called a *compact* for short. A *$\sigma$-compact* set is a countable union of compacts.

A collection $\mathcal{A}$ of subsets of a set $X$ is a *$\sigma$-field in $X$* (sometimes called a *$\sigma$-algebra*) if:

(i) $X \in \mathcal{A}$;

(ii) If $U \in \mathcal{A}$, then $X - U \in \mathcal{A}$;

(iii) The countable union $\bigcup_{j=1}^{\infty} U_j \in \mathcal{A}$ whenever $U_j \in \mathcal{A}$ for all $j \geq 1$.

Note that (iii) clearly includes finite unions. When (iii) is only required to hold for finite unions, then $\mathcal{A}$ is called a *field*. When $\mathcal{A}$ is a $\sigma$-field in $X$, then $X$ (or the pair $(X, \mathcal{A})$) is a *measurable space*, and the members of $\mathcal{A}$ are called the *measurable sets* in $X$. If $X$ is a measurable space and $Y$

is a topological space, then a map $f : X \mapsto Y$ is *measurable* if $f^{-1}(U)$ is measurable in $X$ whenever $U$ is open in $Y$.

If $\mathcal{O}$ is a collection of subsets of $X$ (not necessary open), then there exists a smallest $\sigma$-field $\mathcal{A}^*$ in $X$ so that $\mathcal{O} \in \mathcal{A}^*$. This $\mathcal{A}^*$ is called the $\sigma$-field *generated* by $\mathcal{O}$. To see that such an $\mathcal{A}^*$ exists, let $\mathcal{S}$ be the collection of all $\sigma$-fields in $X$ which contain $\mathcal{O}$. Since the collection of all subsets of $X$ is one such $\sigma$-field, $\mathcal{S}$ is not empty. Define $\mathcal{A}^*$ to be the intersection of all $\mathcal{A} \in \mathcal{S}$. Clearly, $\mathcal{O} \in \mathcal{A}^*$ and $\mathcal{A}^*$ is in every $\sigma$-field containing $\mathcal{O}$. All that remains is to show that $\mathcal{A}^*$ is itself a $\sigma$-field. Assume that $A_j \in \mathcal{A}^*$ for all integers $j \geq 1$. If $\mathcal{A} \in \mathcal{S}$, then $\bigcup_{j\geq 1} A_j \in \mathcal{A}$. Since $\bigcup_{j\geq 1} A_j \in \mathcal{A}$ for every $\mathcal{A} \in \mathcal{S}$, we have $\bigcup_{j\geq 1} A_j \in \mathcal{A}^*$. Also $X \in \mathcal{A}^*$ since $X \in \mathcal{A}$ for all $\mathcal{A} \in \mathcal{S}$; and for any $A \in \mathcal{A}^*$, both $A$ and $X - A$ are in every $\mathcal{A} \in \mathcal{S}$. Thus $\mathcal{A}^*$ is indeed a $\sigma$-field.

A $\sigma$-field is *separable* if it is generated by a countable collection of subsets. Note that we have already defined "separable" as a characteristic of certain topological spaces. There is a connection between the two definitions which we will point out in a few pages during our discussion on metric spaces. When $X$ is a topological space, the smallest $\sigma$-field $\mathcal{B}$ generated by the open sets is called the *Borel $\sigma$-field* of $X$. Elements of $\mathcal{B}$ are called *Borel sets*. A function $f : X \mapsto Y$ between topological spaces is *Borel-measurable* if it is measurable with respect to the Borel $\sigma$-field of $X$. Clearly, a continuous function between topological spaces is also Borel-measurable.

For a $\sigma$-field $\mathcal{A}$ in a set $X$, a map $\mu : \mathcal{A} \mapsto \bar{\mathbb{R}}$ is a *measure* if:

(i) $\mu(A) \in [0, \infty]$ for all $A \in \mathcal{A}$;

(ii) $\mu(\emptyset) = 0$;

(iii) For a disjoint sequence $\{A_j\} \in \mathcal{A}$, $\mu\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mu(A_j)$ (*countable additivity*).

If $X = A_1 \cup A_2 \cup \cdots$ for some finite or countable sequence of sets in $\mathcal{A}$ with $\mu(A_j) < \infty$ for all indices $j$, then $\mu$ is *$\sigma$-finite*. The triple $(X, \mathcal{A}, \mu)$ is called a *measure space*. If $\mu(X) = 1$, then $\mu$ is a *probability measure*. For a probability measure $P$ on a set $\Omega$ with $\sigma$-field $\mathcal{A}$, the triple $(\Omega, \mathcal{A}, P)$ is called a *probability space*. If the set $[0, \infty]$ in Part (i) is extended to $(-\infty, \infty]$ or replaced by $[-\infty, \infty)$ (but not both), then $\mu$ is a *signed measure*. For a measure space $(X, \mathcal{A}, \mu)$, let $\mathcal{A}^*$ be the collection of all $E \subset X$ for which there exists $A, B \in \mathcal{A}$ with $A \subset E \subset B$ and $\mu(B - A) = 0$, and define $\mu(E) = \mu(A)$ in this setting. Then $\mathcal{A}^*$ is a $\sigma$-field, $\mu$ is still a measure, and $\mathcal{A}^*$ is called the *$\mu$-completion* of $\mathcal{A}$.

A *metric space* is a set $\mathbb{D}$ together with a *metric*. A metric or *distance function* is a map $d : \mathbb{D} \times \mathbb{D} \mapsto [0, \infty)$ where:

(i) $d(x, y) = d(y, x)$;

(ii) $d(x, z) \leq d(x, y) + d(y, z)$ (the *triangle inequality*);

(iii) $d(x, y) = 0$ if and only if $x = y$.

A *semimetric* or *pseudometric* satisfies (i) and (ii) but not necessarily (iii). Technically, a metric space consists of the pair $(\mathbb{D}, d)$, but usually only $\mathbb{D}$ is given and the underlying metric $d$ is implied by the context. This is similar to topological and measurable spaces, where only the set of all points $X$ is given while the remaining components are omited except where needed to clarify the context. A semimetric space is also a topological space with the open sets generated by applying arbitrary unions to the *open r-balls* $B_r(x) \equiv \{y : d(x, y) < r\}$ for $r \geq 0$ and $x \in \mathbb{D}$ (where $B_0(x) \equiv \emptyset$). A metric space is also Hausdorff, and, in this case, a sequence $\{x_n\} \in \mathbb{D}$ converges to $x \in \mathbb{D}$ if $d(x_n, x) \to 0$. For a semimetric space, $d(x_n, x) \to 0$ ensures only that $x_n$ converges to elements in the *equivalence class* of $x$, where the equivalence class of $x$ consists of all $\{y \in \mathbb{D} : d(x, y) = 0\}$. Accordingly, the closure $\overline{A}$ of a set $A \in \mathbb{D}$ is not only the smallest closed set containing $A$, as stated earlier, but $\overline{A}$ also equals the set of all points that are limits of sequences $\{x_n\} \in A$. Showing this relationship is saved as an exercise. In addition, two semimetrics $d_1$ and $d_2$ on a set $\mathbb{D}$ are considered equivalent (in a topological sense) if they both generate the same open sets. It is left as an exercise to show that equivalent metrics yield the same convergent subsequences.

A map $f : \mathbb{D} \mapsto \mathbb{E}$ between two semimetric spaces is *continuous at a point* $x$ if and only if $f(x_n) \to f(x)$ for every sequence $x_n \to x$. The map $f$ is continuous (in the topological sense) if and only if it is continuous at all points $x \in \mathbb{D}$. Verifying this last equivalence is saved as an exercise. The following lemma helps to define *semicontinuity* for real valued maps:

LEMMA 6.1 *Let $f : \mathbb{D} \mapsto \mathbb{R}$ be a function on the metric space $\mathbb{D}$. Then the following are equivalent:*

*(i) For all $c \in \mathbb{R}$, the set $\{y : f(y) \geq c\}$ is closed.*

*(ii) For all $y_0 \in \mathbb{D}$, $\limsup_{y \to y_0} f(y) \leq f(y_0)$.*

**Proof.** Assume (i) holds but that (ii) is untrue from some $y_0 \in \mathbb{D}$. This implies that for some $\delta > 0$, $\limsup_{y \to y_0} f(y) = f(y_0) + \delta$. Thus $H \cap \{y : d(y, y_0) < \epsilon\}$, where $H \equiv \{y : f(y) \geq f(y_0) + \delta\}$, is nonempty for all $\epsilon > 0$. Since $H$ is closed by (i), we now have that $y_0 \in H$. But this implies that $f(y_0) = f(y_0) + \delta$, which is impossible. Hence (ii) holds. The proof that (ii) implies (i) is saved as an exercise.□

A function $f : \mathbb{D} \mapsto \mathbb{R}$ satisfying either (i) or (ii) (and hence both) of the conditions in Lemma 6.1 is said to be *upper semicontinuous*. A function $f : \mathbb{D} \mapsto \mathbb{R}$ is *lower semicontinuous* if $-f$ is upper semicontinuous. Using Condition (ii), it is easy to see that a function which is both upper and lower semicontinuous is also continuous. The set of all continuous and bounded functions $f : \mathbb{D} \mapsto \mathbb{R}$, which we denote $C_b(\mathbb{D})$, plays an important role in weak convergence on the metric space $\mathbb{D}$, which we will explore in

Chapter 7. It is not hard to see that the Borel $\sigma$-field on a metric space $\mathbb{D}$ is the smallest $\sigma$-field generated by the open balls. It turns out that the Borel $\sigma$-field $\mathcal{B}$ of $\mathbb{D}$ is also the smallest $\sigma$-field $\mathcal{A}$ making all of $C_b(\mathbb{D})$ measurable. To see this, note that any closed $A \subset \mathbb{D}$ is the preimage of the closed set $\{0\}$ for the continuous bounded function $x \mapsto d(x, A) \wedge 1$, where for any set $B \subset \mathbb{D}$, $d(x, B) \equiv \inf\{d(x, y) : y \in B\}$. Thus $\mathcal{B} \subset \mathcal{A}$. Since it is obvious that $\mathcal{A} \subset \mathcal{B}$, we now have $\mathcal{A} = \mathcal{B}$. A Borel-measurable map $X : \Omega \mapsto \mathbb{D}$ defined on a probability space $(\Omega, \mathcal{A}, P)$ is called a *random element* or *random map* with values in $\mathbb{D}$. Borel measurability is, in many ways, the natural concept to use on metric spaces since it connects nicely with the topological structure.

A *Cauchy sequence* is a sequence $\{x_n\}$ in a semimetric space $(\mathbb{D}, d)$ such that $d(x_n, x_m) \to 0$ as $n, m \to \infty$. A semimetric space $\mathbb{D}$ is *complete* if every Cauchy sequence has a limit $x \in \mathbb{D}$. Every metric space $\mathbb{D}$ has a completion $\overline{\mathbb{D}}$ which has a dense subset *isometric* with $\mathbb{D}$. Two metric spaces are isometric if there exists a *bijection* (a one-to-one and onto map) between them which preserves distances.

When a metric space $\mathbb{D}$ is separable, and therefore has a countable dense subset, the Borel $\sigma$-field for $\mathbb{D}$ is itself a separable $\sigma$-field. To see this, let $A \in \mathbb{D}$ be a countable dense subset and consider the collection of open balls with centers at points in $A$ and with rational radii. Clearly, the set of such balls is countable and generates all open sets in $\mathbb{D}$. A topological space $X$ is *Polish* if it is separable and if there exists a metric making $X$ into a complete metric space. Hence any complete and separable metric space is Polish. Furthermore, any open subset of a Polish space is also Polish. Examples of Polish spaces include Euclidean spaces and many other interesting spaces that we will explore shortly. A *Suslin set* is the continuous image of a Polish space. If a Suslin set is also a Hausdorff topological space, then it is a *Suslin space*. An *analytic set* is a subset $A$ of a Polish space $(X, \mathcal{O})$, which is Suslin with respect to the relative topology $\{A \cap B : B \in \mathcal{O}\}$. Since there always exists a continuous and onto map $f : X \mapsto A$ for any Borel subset $A$ of a Polish space $(X, \mathcal{O})$, every Borel subset of a Polish space is Suslin and therefore also analytic.

A subset $K$ is *totally bounded* if and only if for every $r > 0$, $K$ can be covered by finitely many open $r$-balls. Furthermore, it can be shown that a subset $K$ of a complete semimetric space is compact if and only if it is totally bounded and closed. A totally bounded subset $K$ is also called *precompact* because every sequence in $K$ has a Cauchy subsequence. To see this, assume $K$ is totally bounded and choose any sequence $\{x_n\} \in K$. There exists a nested series of subsequence indices $\{N_m\}$ and a nested series of $2^{-m}$-balls $\{A_m\} \subset K$, such that for each integer $m \geq 1$, $N_m$ is infinite, $N_{m+1} \subset N_m$, $A_{m+1} \subset A_m$, and $x_j \in A_m$ for all $j \in N_m$. This follows from the total boundedness properties. For each $m \geq 1$, choose a $n_m \in N_m$, and note that the subsequence $\{x_{n_m}\}$ is Cauchy. Now assume every sequence in $K$ has a Cauchy subsequence. It is not difficult to verify

that if $K$ were not totally bounded, then it is possible to come up with a sequence which has no Cauchy subsequences (see Exercise 6.5.4). This relationship between compactness and total boundedness implies that a $\sigma$-compact set in a metric space is separable. These definitions of compactness agree with the previously given compactness properties for Hausdorff spaces. This happens because a semimetric space $\mathbb{D}$ can be made into a metric—and hence Hausdorff—space $\mathbb{D}_H$ by equating points in $\mathbb{D}_H$ with equivalence classes in $\mathbb{D}$.

A very important example of a metric space is a *normed space*. A normed space $\mathbb{D}$ is a vector space (also called a linear space) equipped with a *norm*, and a norm is a map $\|\cdot\| : \mathbb{D} \mapsto [0, \infty)$ such that, for all $x, y \in \mathbb{D}$ and $\alpha \in \mathbb{R}$,

(i) $\|x + y\| \leq \|x\| + \|y\|$ (another triangle inequality);

(ii) $\|\alpha x\| = |\alpha| \times \|x\|$;

(iii) $\|x\| = 0$ if and only if $x = 0$.

A *seminorm* satisfies (i) and (ii) but not necessarily (iii). A normed space is a metric space (and a seminormed space is a semimetric space) with $d(x, y) = \|x - y\|$, for all $x, y \in \mathbb{D}$. A complete normed space is called a *Banach space*. Two seminorms $\|\cdot\|_1$ and $\|\cdot\|_2$ on a set $\mathbb{D}$ are equivalent if the following is true for all $x, \{x_n\} \in \mathbb{D}$: $\|x_n - x\|_1 \to 0$ if and only if $\|x_n - x\|_2 \to 0$.

In our definition of a normed space $\mathbb{D}$, we require the space to also be a vector space (and therefore it contains all linear combinations of elements in $\mathbb{D}$). However, it is sometimes of interest to apply norms to subsets $K \subset \mathbb{D}$ which may not be linear subspaces. In this setting, let $\operatorname{lin} K$ denote the *linear span of* $K$ (all linear combinations of elements in $K$), and let $\overline{\operatorname{lin}}\, K$ the closure of $\operatorname{lin} K$. Note that both $\operatorname{lin} K$ and $\overline{\operatorname{lin}} K$ are now vector spaces and that $\overline{\operatorname{lin}} K$ is also a Banach space.

We now present several specific examples of metric spaces. The Euclidean space $\mathbb{R}^d$ is a Banach space with squared norm $\|x\|^2 = \sum_{j=1}^d x_j^2$. This space is equivalent under several other norms, including $\|x\| = \max_{1 \leq j \leq d} |x_j|$ and $\|x\| = \sum_{j=1}^d |x_j|$. A Euclidean space is separable with a countably dense subset consisting of all vectors with rational coordinates. By the Heine-Borel theorem, a subset in a Euclidean space is compact if and only if it is closed and bounded. The Borel $\sigma$-field is generated by the intervals of the type $(-\infty, x]$, for rational $x$, where the interval is defined as follows: $y \in (-\infty, x]$ if and only if $y_j \in (-\infty, x_j]$ for all coordinates $j = 1, \ldots, d$. For one-dimensional Euclidean space, $\mathbb{R}$, the norm is $\|x\| = |x|$ (absolute value). The extended real line $\bar{\mathbb{R}} = [-\infty, \infty]$ is a metric space with respect to the metric $d(x, y) = |G(x) - G(y)|$, where $G : \bar{\mathbb{R}} \mapsto \mathbb{R}$ is any strictly monotone increasing, continuous and bounded function, such as the arctan function. For any sequence $\{x_n\} \in \bar{\mathbb{R}}$, $|x_n - x| \to 0$ implies $d(x_n, x) \to 0$,

while divergence of $d(x_n, x)$ implies divergence of $|x_n - x|$. In addition, it is possible for a sequence to converge, with respect to $d$, to either $-\infty$ or $\infty$. This makes $\bar{\mathbb{R}}$ compact.

Another important example is the set of bounded real functions $f : T \mapsto \mathbb{R}$, where $T$ is an arbitrary set. This is a vector space if sums $z_1 + z_2$ and products with scalars, $\alpha z$, are defined pointwise for all $z, z_1, z_2 \in \ell^\infty(T)$. Specifically, $(z_1 + z_2)(t) = z_1(t) + z_2(t)$ and $(\alpha z)(t) = \alpha z(t)$, for all $t \in T$. This space is denoted $\ell^\infty(T)$. The *uniform norm* $\|x\|_T \equiv \sup_{t \in T} |x(t)|$ makes $\ell^\infty(T)$ into a Banach space consisting exactly of all functions $z : T \mapsto \mathbb{R}$ satisfying $\|z\|_T < \infty$. It is not hard to show that $\ell^\infty(T)$ is separable if and only if $T$ is countable.

Two useful subspaces of $\ell^\infty([a, b])$, where $a, b \in \bar{\mathbb{R}}$, are $C[a, b]$ and $D[a, b]$. The space $C[a, b]$ consists of continuous functions $z : [a, b] \mapsto \mathbb{R}$, and $D[a, b]$ is the space of *cadlag* functions which are right-continuous with left-hand limits (cadlag is an abbreviation for continue à droite, limites à gauche). We usually equip these spaces with the uniform norm $\|\cdot\|_{[a,b]}$ inherited from $\ell^\infty([a, b])$. Note that $C[a, b] \subset D[a, b] \subset \ell^\infty([a, b])$. Relative to the uniform norm, $C[a, b]$ is separable, and thus also Polish by the completeness established in Exercise 6.5.5(a), but $D[a, b]$ is not separable. Sometimes, $D[a, b]$ is called the *Skorohod space*, although Skorohod equipped $D[a, b]$ with a special metric—quite different than the uniform metric—resulting in a separable space.

An important subspace of $\ell^\infty(T)$ is the space $UC(T, \rho)$, where $\rho$ is a semimetric on $T$. $UC(T, \rho)$ consists of all bounded function $f : T \mapsto \mathbb{R}$ which are uniformly $\rho$-continuous, i.e.,

$$\lim_{\delta \downarrow 0} \sup_{\rho(s,t) < \delta} |f(s) - f(t)| = 0.$$

When $(T, \rho)$ is totally bounded, the boundedness requirement for functions in $UC(T, \rho)$ is superfluous since a uniformly continuous function on a totally bounded set must necessarily be bounded. We denote $C(T, \rho)$ to be the space of $\rho$-continuous (not necessarily continuous) function on $T$. It is left as an exercise to show that the spaces $C[a, b]$, $D[a, b]$, $UC(T, \rho)$, $C(\overline{T}, \rho)$, when $(T, \rho)$ is a totally bounded semimetric space, and $UC(T, \rho)$ and $\ell^\infty(T)$, for an arbitrary set $T$, are all complete with respect to the uniform metric. When $(T, \rho)$ is a compact semimetric space, $T$ is totally bounded, and a $\rho$-continuous function in $T$ is automatically uniformly $\rho$-continuous. Thus, when $T$ is compact, $C(T, \rho) = UC(T, \rho)$. Actually, every space $UC(T, \rho)$ is equivalent to a space $C(\overline{T}, \rho)$, because the completion $\overline{T}$ of a totally bounded space $T$ is compact and, furthermore, every uniformly continuous function on $T$ has a unique continuous extension to $\overline{T}$. Showing this is saved as an exercise.

The foregoing structure makes it clear that $UC(T, \rho)$ is a Polish space that is made complete by the uniform norm. Hence $UC(T, \rho)$ is also $\sigma$-compact. In fact, any $\sigma$-compact set in $\ell^\infty(T)$ is contained in $UC(T, \rho)$,

for some totally bounded semimetric space $(T, \rho)$, and all compact sets in $\ell^\infty(T)$ have a specific form:

THEOREM 6.2 *(Arzelà-Ascoli)*

*(a) The closure of $K \subset UC(T, \rho)$, where $(T, \rho)$ is totally bounded, is compact if and only if*

*(i)* $\sup_{x \in K} |x(t_0)| < \infty$, *for some* $t_0 \in T$; *and*

*(ii)*

$$\lim_{\delta \downarrow 0} \sup_{x \in K} \sup_{s,t \in T: \rho(s,t) < \delta} |x(s) - x(t)| = 0.$$

*(b) The closure of $K \subset \ell^\infty(T)$ is $\sigma$-compact if and only if $K \subset UC(T, \rho)$ for some semimetric $\rho$ making $T$ totally bounded.*

The proof is given in Section 6.4. Since all compact sets are trivially $\sigma$-compact, Theorem 6.2 implies that any compact set in $\ell^\infty(T)$ is actually contained in $UC(T, \rho)$ for some semimetric $\rho$ making $T$ totally bounded.

Another important class of metric spaces are product spaces. For a pair of metric spaces $(\mathbb{D}, d)$ and $(\mathbb{E}, e)$, the *Cartesian product* $\mathbb{D} \times \mathbb{E}$ is a metric space with respect to the metric $\rho((x_1, y_1), (x_2, y_2)) \equiv d(x_1, x_2) \vee e(y_1, y_2)$, for $x_1, x_2 \in \mathbb{D}$ and $y_1, y_2 \in \mathbb{E}$. This resulting topology is the *product topology*. In this setting, convergence of $(x_n, y_n) \to (x, y)$ is equivalent to convergence of both $x_n \to x$ and $y_n \to y$. There are two natural $\sigma$-fields for $\mathbb{D} \times \mathbb{E}$ that we can consider. The first is the Borel $\sigma$-field for $\mathbb{D} \times \mathbb{E}$ generated from the product topology. The second is the product $\sigma$-field generated by all sets of the form $A \times B$, where $A \in \mathcal{A}$, $B \in \mathcal{B}$, and $\mathcal{A}$ and $\mathcal{B}$ are the respective $\sigma$-fields for $\mathbb{D}$ and $\mathbb{E}$. These two are equal when $\mathbb{D}$ and $\mathbb{E}$ are separable, but they may be unequal otherwise, with the first $\sigma$-field larger than the second. Suppose $X : \Omega \mapsto \mathbb{D}$ and $Y : \Omega \mapsto \mathbb{E}$ are Borel-measurable maps defined on a measurable space $(\Omega, \mathcal{A})$. Then $(X, Y) : \Omega \mapsto \mathbb{D} \times \mathbb{E}$ is a measurable map for the product of the two $\sigma$-fields by the definition of a measurable map. Unfortunately, when the Borel $\sigma$-field for $\mathbb{D} \times \mathbb{E}$ is larger than the product $\sigma$-field, then it is possible for $(X, Y)$ to not be Borel-measurable.

## 6.2 Outer Expectation

An excellent overview of outer expectations is given in Chapter 1.2 of van der Vaart and Wellner (1996). The concept applies to an arbitrary probability space $(\Omega, \mathcal{A}, P)$ and an arbitrary map $T : \Omega \mapsto \bar{\mathbb{R}}$, where $\bar{\mathbb{R}} \equiv [-\infty, \infty]$. As described in Chapter 2, the outer expectation of $T$, denoted $\mathrm{E}^* T$, is the infimum over all $\mathrm{E}U$, where $U : \Omega \mapsto \bar{\mathbb{R}}$ is measurable, $U \geq T$, and $\mathrm{E}U$ exists. For $\mathrm{E}U$ to exist, it must not be indeterminate, although it can be

$\pm\infty$, provided the sign is clear. Since $T$ is not necessarily a random variable, the proper term for $\mathrm{E}^*T$ is *outer integral*. However, we will use the term outer expectation throughout the remainder of this book in deference to its connection with the classical notion of expectation. We analogously define inner expectation: $\mathrm{E}_*T = -\mathrm{E}^*[-T]$. The following lemma verifies the existence of a minimal measurable majorant $T^* \geq T$:

LEMMA 6.3 *For any $T : \Omega \mapsto \bar{\mathbb{R}}$, there exists a minimal measurable majorant $T^* : \Omega \mapsto \bar{\mathbb{R}}$ with*

*(i) $T^* \geq T$;*

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

*For any $T^*$ satisfying (i) and (ii), $\mathrm{E}^*T = \mathrm{E}T^*$, provided $\mathrm{E}T^*$ exists. The last statement is true if $\mathrm{E}^*T < \infty$.*

The proof is given in Section 6.4 at the end of this chapter. The following lemma, the proof of which is left as an exercise, is an immediate consequence of Lemma 6.3 and verifies the existence of a maximal measurable minorant:

LEMMA 6.4 *For any $T : \Omega \mapsto \bar{\mathbb{R}}$, the maximal measurable minorant $T_* \equiv -(-T)^*$ exists and satisfies*

*(i) $T_* \leq T$;*

*(ii) For every measurable $U : \Omega \mapsto \bar{\mathbb{R}}$ with $U \leq T$ a.s., $T_* \geq U$ a.s.*

*For any $T_*$ satisfying (i) and (ii), $\mathrm{E}_*T = \mathrm{E}T_*$, provided $\mathrm{E}T_*$ exists. The last statement is true if $\mathrm{E}_*T > -\infty$.*

An important special case of outer expectation is outer probability. The outer probability of an arbitrary $B \subset \Omega$, denoted $P^*(B)$, is the infimum over all $P(A)$ such that $A \supset B$ and $A \in \mathcal{A}$. The inner probability of an arbitrary $B \subset \Omega$ is defined to be $P_*(B) = 1 - P^*(\Omega - B)$. The following lemma gives the precise connection between outer/inner expectations and outer/inner probabilities:

LEMMA 6.5 *For any $B \subset \Omega$,*

*(i) $P^*(B) = \mathrm{E}^*1\{B\}$ and $P_*(B) = \mathrm{E}_*1\{B\}$;*

*(ii) there exists a measurable set $B^* \supset B$ so that $P(B^*) = P^*(B)$; for any such $B^*$, $1\{B^*\} = (1\{B\})^*$;*

*(iii) For $B_* \equiv \Omega - \{\Omega - B\}^*$, $P_*(B) = P(B_*)$;*

*(iv) $(1\{B\})^* + (1\{\Omega - B\})_* = 1$.*

**Proof.** From the definitions, $P^*(B) = \inf_{\{A \in \mathcal{A}: A \supset B\}} \mathrm{E}1\{A\} \geq \mathrm{E}^*1\{B\}$. Next, $\mathrm{E}^*1\{B\} = \mathrm{E}(1\{B\})^* \geq \mathrm{E}1\{(1\{B\})^* \geq 1\} = P\{(1\{B\})^* \geq 1\} \geq$

$P^*(B)$, where the last inequality follows from the definition of $P^*$. Combining the two conclusions yields that all inequalities are actually equalities. This gives the first parts of (i) and (ii), with $B^* = \{(1\{B\})^* \geq 1\}$. The second part of (i) results from $P_*(B) = 1 - P^*(\Omega - B) = 1 - \mathrm{E}(1 - 1\{B\})^* = 1 - \mathrm{E}\,(1 - (1\{B\})_*)$. The second part of (ii) follows from $(1\{B\})^* \leq 1\{B^*\} = 1\,\{(1\{B\})^* \geq 1\} \leq (1\{B\})^*$. The definition of $P_*$ implies (iii) directly. To verify (iv), we have $(1\{\Omega - B\})_* = (1 - 1\{B\})_* = -(1\{B\} - 1)^* = 1 - (1\{B\})^*$.□

The following three lemmas, Lemmas 6.6–6.8, provide several relations which will prove useful later on but which might be skipped on a first reading. The proofs are given in Section 6.4 and in the exercises.

LEMMA 6.6 *Let $S, T : \Omega \mapsto \mathbb{R}$ be arbitrary maps. The following statements are true almost surely, provided the statements are well-defined:*

*(i)* $S_* + T^* \leq (S+T)^* \leq S^* + T^*$, *with all equalities if $S$ is measurable;*

*(ii)* $S_* + T_* \leq (S+T)_* \leq S_* + T^*$, *with all equalities if $T$ is measurable;*

*(iii)* $(S-T)^* \geq S^* - T^*$;

*(iv)* $|S^* - T^*| \leq |S-T|^*$;

*(v)* $(1\{T > c\})^* = 1\{T^* > c\}$, *for any $c \in \mathbb{R}$;*

*(vi)* $(1\{T \geq c\})_* = 1\{T_* \geq c\}$, *for any $c \in \mathbb{R}$;*

*(vii)* $(S \vee T)^* = S^* \vee T^*$;

*(viii)* $(S \wedge T)^* \leq S^* \wedge T^*$, *with equality if $S$ is measurable.*

LEMMA 6.7 *For any sets $A, B \subset \Omega$,*

*(i)* $(A \cup B)^* = A^* \cup B^*$ *and* $(A \cap B)_* = A_* \cap B_*$;

*(ii)* $(A \cap B)^* \subset A^* \cap B^*$ *and* $(A \cup B)_* \supset A_* \cup B_*$, *with the inclusions replaced by equalities if either $A$ or $B$ is measurable;*

*(iii) If* $A \cap B = \emptyset$, *then* $P_*(A) + P_*(B) \leq P_*(A \cup B) \leq P^*(A \cup B) \leq P^*(A) + P^*(B)$.

LEMMA 6.8 *Let $T : \Omega \mapsto \mathbb{R}$ be an arbitrary map and let $\phi : \mathbb{R} \mapsto \mathbb{R}$ be monotone, with an extension to $\bar{\mathbb{R}}$. The following statements are true almost surely, provided the statements are well-defined:*

*A. If $\phi$ is nondecreasing, then*

*(i)* $\phi(T^*) \geq [\phi(T)]^*$, *with equality if $\phi$ is left-continuous on $[-\infty, \infty)$;*

*(ii)* $\phi(T_*) \leq [\phi(T)]_*$, *with equality if $\phi$ is right-continuous on $(-\infty, \infty]$.*

*B. If $\phi$ is nonincreasing, then*

*(i)* $\phi(T^*) \le [\phi(T)]_*$, *with equality if* $\phi$ *is left-continuous on* $[-\infty, \infty)$;

*(ii)* $\phi(T_*) \ge [\phi(T)]^*$, *with equality if* $\phi$ *is right-continuous on* $(-\infty, \infty]$.

We next present an outer-expectation version of the unconditional Jensen's inequality:

LEMMA 6.9 **(Jensen's inequality)** *Let* $T : \Omega \mapsto \mathbb{R}$ *be an arbitrary map, with* $\mathrm{E}^*|T| < \infty$, *and assume* $\phi : \mathbb{R} \mapsto \mathbb{R}$ *is convex. Then*

*(i)* $\mathrm{E}^*\phi(T) \ge \phi(\mathrm{E}_*T) \vee \phi(\mathrm{E}^*T)$;

*(ii) if* $\phi$ *is also monotone,* $\mathrm{E}_*\phi(T) \ge \phi(\mathrm{E}_*T) \wedge \phi(\mathrm{E}^*T)$.

**Proof.** Assume first that $\phi$ is monotone increasing. Since $\phi$ is also continuous (by convexity), $\mathrm{E}^*\phi(T) = \mathrm{E}\phi(T^*) \ge \phi(\mathrm{E}^*T)$, where the equality follows from A(i) of Lemma 6.8 and the inequality from the usual Jensen's inequality. Similar arguments verify that $\mathrm{E}_*\phi(T) \ge \phi(\mathrm{E}_*T)$ based on A(ii) of the same lemma. Note also that $\phi(\mathrm{E}^*T) \ge \phi(\mathrm{E}_*T)$. Now assume that $\phi$ is monotone decreasing. Using B(i) and B(ii) of Lemma 6.8 and arguments similar to those used for increasing $\phi$, we obtain that both $\mathrm{E}^*\phi(T) \ge \phi(\mathrm{E}_*T)$ and $\mathrm{E}_*\phi(T) \ge \phi(\mathrm{E}^*T)$. Note in this case that $\phi(\mathrm{E}_*T) \ge \phi(\mathrm{E}^*T)$. Thus when $\phi$ is monotone (either increasing or decreasing), we have that both $\mathrm{E}^*\phi(T) \ge \phi(\mathrm{E}^*T) \vee \phi(\mathrm{E}_*T)$ and $\mathrm{E}_*\phi(T) \ge \phi(\mathrm{E}^*T) \wedge \phi(\mathrm{E}_*T)$. Hence (ii) follows.

We have also proved (i) in the case where $\phi$ is monotone. Suppose now that $\phi$ is not monotone. Then there exists a $c \in \mathbb{R}$ so that $\phi$ is nonincreasing over $(-\infty, c]$ and nondecreasing over $(c, \infty)$. Let $g_1(t) \equiv \phi(t)1\{t \le c\} + \phi(c)1\{t > c\}$ and $g_2(t) \equiv \phi(c)1\{t \le c\} + \phi(t)1\{t > c\}$, and note that $\phi(t) = g_1(t) + g_2(t) - \phi(c)$ and that both $g_1$ and $g_2$ are convex and monotone. Now $[\phi(T)]^* = [g_1(T) + g_2(T) - \phi(c)]^* \ge [g_1(T)]_* + [g_2(T)]^* - \phi(c)$ by Part (i) of Lemma 6.6. Now, using the results in the previous paragraph, we have that $\mathrm{E}^*\phi(T) \ge g_1(\mathrm{E}^*T) + g_2(\mathrm{E}^*T) - \phi(c) = \phi(\mathrm{E}^*T)$. However, we also have that $[g_1(T) + g_2(T) - \phi(c)]^* \ge [g_1(T)]^* + [g_2(T)]_* - \phi(c)$. Thus, again using results from the previous paragraph, we have $\mathrm{E}^*\phi(T) \ge g_1(\mathrm{E}_*T) + g_2(\mathrm{E}_*T) - \phi(c) = \phi(\mathrm{E}_*T)$. Hence (i) follows.□

The following outer-expectation version of Chebyshev's inequality is also useful:

LEMMA 6.10 **(Chebyshev's inequality)** *Let* $T : \Omega \mapsto \mathbb{R}$ *be an arbitrary map, with* $\phi : [0, \infty) \mapsto [0, \infty)$ *nondecreasing and strictly positive on* $(0, \infty)$. *Then, for every* $u > 0$,

$$P^*(|T| \ge u) \le \frac{\mathrm{E}^*\phi(|T|)}{\phi(u)}.$$

**Proof.** The result follows from the standard Chebyshev inequality as a result of the following chain of inequalities:

$$(1\{|T| \geq u\})^* \leq (1\{\phi(|T|) \geq \phi(u)\})^* \leq 1\left\{[\phi(|T|)]^* \geq \phi(u)\right\}.$$

The first inequality follows from the fact that $|T| \geq u$ implies $\phi(|T|) \geq \phi(u)$, and the second inequality follows from A(i) of Lemma 6.8.□

There are many other analogies between outer and standard versions of expectation and probability, but we only present a few more in this chapter. We next present versions of the monotone and dominated convergence theorems. The proofs are given in Section 6.4. Some additional results are given in the exercises.

LEMMA 6.11 **(Monotone convergence)** *Let $T_n, T : \Omega \mapsto \mathbb{R}$ be arbitrary maps on a probability space, with $T_n \uparrow T$ pointwise on a set of inner probability one. Then $T_n^* \uparrow T^*$ almost surely. Provided $\mathrm{E}^* T_n > -\infty$ for some $n$, then $\mathrm{E}^* T_n \uparrow \mathrm{E}^* T$.*

LEMMA 6.12 **(Dominated convergence)** *Let $T_n, T, S : \Omega \mapsto \mathbb{R}$ be maps on a probability space, with $|T_n - T|^* \xrightarrow{\text{as}} 0$, $|T_n| \leq S$ for all $n$, and $\mathrm{E}^* S < \infty$. Then $\mathrm{E}^* T_n \to \mathrm{E}^* T$.*

Let $(\Omega, \tilde{\mathcal{A}}, \tilde{P})$ be the $P$-completion of the probability space $(\Omega, \mathcal{A}, P)$, as defined in the previous section (for general measure spaces). Recall that a completion of a measure space is itself a measure space. One can also show that $\tilde{\mathcal{A}}$ is the $\sigma$-field of all sets of the form $A \cup N$, with $A \in \mathcal{A}$ and $N \subset \Omega$ so that $P^*(N) = 0$, and $\tilde{P}$ is the probability measure satisfying $\tilde{P}(A \cup N) = P(A)$. A nice property of $\left(\Omega, \tilde{\mathcal{A}}, \tilde{P}\right)$ is that for every measurable map $\tilde{S} : (\Omega, \tilde{\mathcal{A}}) \mapsto \mathbb{R}$, there is a measurable map $S : (\Omega, A) \mapsto \mathbb{R}$ such that $P^*(S \neq \tilde{S}) = 0$. Furthermore, a minimal measurable cover $T^*$ of a map $T : (\Omega, \mathcal{A}, P) \mapsto \bar{\mathbb{R}}$ is a version of a minimal measurable cover $\tilde{T}^*$ for $T$ as a map on the $P$-completion of $(\Omega, \mathcal{A}, P)$, i.e., $P^*(T^* \neq \tilde{T}^*) = 0$. While it is not difficult to show this, we do not prove it.

We close this section with two results which have application to product probability spaces. The first result involves *perfect* maps, and the second result is a special formulation of Fubini's theorem. Consider composing a map $T : \Omega \mapsto \mathbb{R}$ with a measurable map $\phi : \tilde{\Omega} \mapsto \Omega$, defined on some probability space, to form $T \circ \phi : \left(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{P}\right) \mapsto \mathbb{R}$, where $\phi : \left(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{P}\right) \mapsto (\Omega, \mathcal{A}, P)$. Denote $T^*$ as the minimal measurable cover of $T$ for $\tilde{P} \circ \phi^{-1}$. It is easy to see that since $T^* \circ \phi \geq T \circ \phi$, we have $(T \circ \phi)^* \leq T^* \circ \phi$. The map $\phi$ is perfect if $(T \circ \phi)^* = T^* \circ \phi$, for every bounded $T : \Omega \mapsto \mathbb{R}$. This property ensures that $P^*(\phi \in A) = \left(\tilde{P} \circ \phi^{-1}\right)^*(A)$ for every set $A \subset \Omega$.

An important example of a perfect map is a coordinate projection in a product probability space. Specifically, let $T$ be a real valued map defined on $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \times \mathcal{A}_2, P_1 \times P_2)$ which only depends on the first coordinate of $\omega = (\omega_1, \omega_2)$. $T^*$ can then be computed by just ignoring $\Omega_2$ and thinking of $T$ as a map on $\Omega_1$. More precisely, suppose $T = T_1 \circ \pi_1$, where $\pi_1$ is

the projection on the first coordinate. The following lemma shows that $T^* = T_1^* \circ \pi_1$, and thus coordinate projections such as $\pi_1$ are perfect. We will see other examples of perfect maps later on in Chapter 7.

LEMMA 6.13 *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 onto the first coordinate, and let $T : \Omega_1 \mapsto \mathbb{R}$ by bounded, but otherwise arbitrary. Let $T^*$ be the minimal measurable cover of $T$ for $P_1 = (P_1 \times P_2) \circ \pi_1^{-1}$. By definition, $(T \circ \pi_1)^* \leq T^* \circ \pi_1$. Now suppose $U \geq T \circ \pi_1$, $P_1 \times P_2$-a.s., and is measurable, where $U : \Omega_1 \times \Omega_2 \mapsto \mathbb{R}$. Fubini's theorem yields that for $P_2$-almost all $\Omega_2$, we have $U(\omega_1, \omega_2) \geq T(\omega_1)$ for $P_2$-almost all $\omega_2$. But for fixed $\omega_2$, $U$ is a measurable function of $\omega_1$. Thus for $P_2$-almost all $\omega_2$, $U(\omega_1, \omega_2) \geq T^*(\omega_1)$ for $P_1$-almost all $\omega_1$. Applying Fubinin's theorem again, the jointly measurable set $\{(\omega_1, \omega_2) : U < T^* \circ \pi_1\}$ is $P_1 \times P_2$-null. Hence $(T \circ \pi_1)^* = T^* \circ \pi_1$ almost surely.□

Now we consider Fubini's theorem for maps on product spaces which may not be measurable. There is no generally satisfactory version of Fubini's theorem that will work in all nonmeasurable settings of interest, and it is frequently necessary to establish at least some kind of measurability to obtain certain key empirical process results. The version of Fubini's theorem we now present basically states that repeated outer expectations are always less than joint outer expectations. Let $T$ be an arbitrary real map defined on the product space $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \times \mathcal{A}_2, P_1 \times P_2)$. We write $\mathrm{E}_1^*\mathrm{E}_2^* T$ to mean outer expectations taken in turn. For fixed $\omega_1$, let $(\mathrm{E}_2^* T)(\omega_1)$ be the infimum of $\int_{\Omega_2} U(\omega_2) dP_2(\omega_2)$ taken over all measurable $U : \Omega_2 \mapsto \bar{\mathbb{R}}$ with $U(\omega_2) \geq T(\omega_1, \omega_2)$ for every $\omega_2$ for which $\int_{\Omega_2} U(\omega_2) dP_2(\omega_2)$ exists. Next, $\mathrm{E}_1^*\mathrm{E}_2^* T$ is the outer integral of $\mathrm{E}_2^* T : \Omega_1 \mapsto \bar{\mathbb{R}}$. Repeated inner expectations are analogously defined. The following version of Fubini's theorem gives bounds for this repeated expectation process. We omit the proof.

LEMMA 6.14 **(Fubini's theorem)** *Let $T$ be an arbitrary real valued map on a product probability space. Then* $\mathrm{E}_* T \leq \mathrm{E}_{1*}\mathrm{E}_{2*} T \leq \mathrm{E}_1^*\mathrm{E}_2^* T \leq \mathrm{E}^* T$.

## 6.3 Linear Operators and Functional Differentiation

A *linear operator* is a map $T : \mathbb{D} \mapsto \mathbb{E}$ between normed spaces with the property that $T(ax + by) = aT(x) + bT(y)$ for all scalars $a, b$ and any $x, y \in \mathbb{D}$. When the range space $\mathbb{E}$ is $\mathbb{R}$, then $T$ is a *linear functional*. When $T$ is linear, we will often use $Tx$ instead of $T(x)$. A linear operator $T : \mathbb{D} \mapsto \mathbb{E}$ is a *bounded linear operator* if

$$\|T\| \equiv \sup_{x\in\mathbb{D}:\|x\|\leq 1} \|Tx\| < \infty. \tag{6.1}$$

Here, the norms $\|\cdot\|$ are defined by the context. We have the following proposition:

PROPOSITION 6.15 *For a linear operator $T : \mathbb{D} \mapsto \mathbb{E}$ between normed spaces, the following are equivalent:*

*(i) $T$ is continuous at a point $x_0 \in \mathbb{D}$;*

*(ii) $T$ is continuous on all of $\mathbb{D}$;*

*(iii) $T$ is bounded.*

**Proof.** We save the implication (i) $\Rightarrow$ (ii) as an exercise. Note that by linearity, boundedness of $T$ is equivalent to there existing some $0 < c < \infty$ for which

$$\|Tx\| \leq c\|x\| \text{ for all } x \in \mathbb{D}. \tag{6.2}$$

Assume $T$ is continuous but that there exists no $0 < c < \infty$ satisfying (6.2). Then there exists a sequence $\{x_n\} \in \mathbb{D}$ so that $\|x_n\| = 1$ and $\|Tx_n\| \geq n$ for all $n \geq 1$. Define $y_n = \|Tx_n\|^{-1}x_n$ and note that $\|Ty_n\| = 1$ by linearity. Now $y_n \to 0$ and thus $Ty_n \to 0$, but this is a contradiction. Thus there exists some $0 < c < \infty$ satisfying (6.2), and (iii) follows. Now assume (iii) and let $\{x_n\} \in X$ be any sequence satisfying $x_n \to 0$. Then by (6.2), $\|Tx_n\| \to 0$, and thus (i) is satisfied at $x_0 = 0$.□

For normed spaces $\mathbb{D}$ and $\mathbb{E}$, let $B(\mathbb{D}, \mathbb{E})$ be the space of all bounded linear operators $T : \mathbb{D} \mapsto \mathbb{E}$. This structure makes the space $B(\mathbb{D}, \mathbb{E})$ into a normed space with norm $\|\cdot\|$ defined in (6.1). When $\mathbb{E}$ is a Banach space, then any convergent sequence $T_nx_n$ will be contained in $\mathbb{E}$, and thus $B(\mathbb{D}, \mathbb{E})$ is also a Banach space. When $\mathbb{D}$ is not a Banach space, $T$ has a unique continuous extension to $\overline{\mathbb{D}}$. To see this, fix $x \in \overline{\mathbb{D}}$ and let $\{x_n\} \in \mathbb{D}$ be a sequence converging to $x$. Then, since $\|Tx_n - Tx_m\| \leq c\|x_n - x_m\|$, $Tx_n$ converges to some point in $\overline{\mathbb{E}}$. Next, note that if both sequences $\{x_n\}, \{y_n\} \in \mathbb{D}$ converge to $x$, then $\|Tx_n - Ty_n\| \leq c\|x_n - y_n\| \to 0$. Thus we can define an extension $\overline{T} : \overline{\mathbb{D}} \mapsto \overline{\mathbb{E}}$ to be the unique linear operator with $\overline{T}x = \lim_{n\to\infty} Tx_n$, where $x$ is any point in $\overline{\mathbb{D}}$ and $\{x_n\}$ is any sequence in $\mathbb{D}$ converging to $x$.

For normed spaces $\mathbb{D}$ and $\mathbb{E}$, and for any $T \in B(\mathbb{D}, \mathbb{E})$, $N(T) \equiv \{x \in \mathbb{D} : Tx = 0\}$ is the *null space* of $T$ and $R(T) \equiv \{y \in \mathbb{E} : Tx = y \text{ for some } x \in \mathbb{D}\}$ is the *range space* of $T$. It is clear that $T$ is one-to-one if and only if $N(T) = \{0\}$. We have the following two results for inverses, which we give without proof:

LEMMA 6.16 *Assume $\mathbb{D}$ and $\mathbb{E}$ are normed spaces and that $T \in B(\mathbb{D}, \mathbb{E})$. Then*

*(i) $T$ has a continuous inverse $T^{-1} : R(T) \mapsto \mathbb{D}$ if and only if there exists a $c > 0$ so that $\|Tx\| \geq c\|x\|$ for all $x \in \mathbb{D}$;*

*(ii)* **(Banach's theorem)** *If $\mathbb{D}$ and $\mathbb{E}$ are complete and $T$ is continuous with $N(T) = \{0\}$, then $T^{-1}$ is continuous if and only if $R(T)$ is closed.*

A linear operator $T : \mathbb{D} \mapsto \mathbb{E}$ between normed spaces is a *compact operator* if $\overline{T(U)}$ is compact in $\overline{\mathbb{E}}$, where $U = \{x \in \mathbb{D} : \|x\| \leq 1\}$ is the unit ball in $\mathbb{D}$ and, for a set $A \in \mathbb{D}$, $T(A) \equiv \{Tx : x \in A\}$. The operator $T$ is *onto* if for every $y \in \mathbb{E}$, there exists an $x \in \mathbb{D}$ so that $Tx = y$. Later on in the book, we will encounter linear operators of the form $T + K$, where $T$ is continuously invertible and onto and $K$ is compact. The following result will be useful:

LEMMA 6.17 *Let $A = T + K : \mathbb{D} \mapsto \mathbb{E}$ be a linear operator between Banach spaces, where $T$ is both continuously invertible and onto and $K$ is compact. Then if $N(A) = \{0\}$, $A$ is also continuously invertible and onto.*

**Proof.** We only sketch the proof. Since $T^{-1}$ is continuous, the operator $T^{-1}K : \mathbb{E} \mapsto \mathbb{D}$ is compact. Hence $I + T^{-1}K$ is one-to-one and therefore also onto by a result of Riesz for compact operators (see, for example, Theorem 3.4 of Kress, 1999). Thus $T + K$ is also onto. We will be done if we can show that $I + T^{-1}K$ is continuously invertible, since that would imply that $(T+K)^{-1} = (I+T^{-1}K)^{-1}T^{-1}$ is bounded. Let $L \equiv I+T^{-1}K$ and assume $L^{-1}$ is not bounded. Then there exists a sequence $\{x_n\} \in \mathbb{D}$ with $\|x_n\| = 1$ and $\|L^{-1}x_n\| \geq n$ for all integers $n \geq 1$. Let $y_n = \left(\|L^{-1}x_n\|\right)^{-1} x_n$ and $\phi_n = \left(\|L^{-1}x_n\|\right)^{-1} L^{-1}x_n$, and note that $y_n \to 0$ while $\|\phi_n\| = 1$ for all $n \geq 1$. Since $T^{-1}K$ is compact, there exists a subsequence $\{n'\}$ so that $T^{-1}K\phi_{n'} \to \phi \in \mathbb{D}$. Since $\phi_n + T^{-1}K\phi_n = y_n$ for all $n \geq 1$, we have $\phi_{n'} \to -\phi$. Hence $\phi \in N(L)$, which implies $\phi = 0$ since $L$ is one-to-one. But this contradicts $\|\phi_n\| = 1$ for all $n \geq 1$. Thus $L^{-1}$ is bounded.□

The following simple inversion result for *contraction operators* is also useful. An operator $A$ is a contraction operator if $\|A\| < 1$.

PROPOSITION 6.18 *Let $A : \mathbb{D} \mapsto \mathbb{D}$ be a linear operator with $\|A\| < 1$. Then $I - A$, where $I$ is the identity, is continuously invertible and onto with inverse $(I - A)^{-1} = \sum_{j=0}^{\infty} A^j$.*

**Proof.** Let $B \equiv \sum_{j=0}^{\infty} A^j$, and note that $\|B\| \leq \sum_{j=0}^{\infty} \|A\|^j = (1 - \|A\|)^{-1} < \infty$. Thus $B$ is a bounded linear operator on $\mathbb{D}$. Since $(I-A)B = I$ by simple algebra, we have that $B = (I - A)^{-1}$, and the result follows.□

We now shift our attention to differentiation. Let $\mathbb{D}$ and $\mathbb{E}$ be two normed spaces, and let $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ be a function. We allow the domain $\mathbb{D}_\phi$ of the function to be an arbitrary subset of $\mathbb{D}$. The function $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ is *Gâteaux-differentiable* at $\theta \in \mathbb{D}_\phi$, in the direction $h$, if there exists a quantity $\phi'_\theta(h) \in \mathbb{E}$ so that

$$\frac{\phi(\theta + t_n h) - \phi(\theta)}{t_n} \to \phi'_\theta(h),$$

as $n \to \infty$, for any scalar sequence $t_n \to 0$. Gâteaux differentiability is usually not strong enough for the applications of functional derivatives needed for Z-estimators and the delta method. The stronger differentiability we need is *Hadamard* and *Fréchet* differentiability.

A map $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ is Hadamard differentiable at $\theta \in \mathbb{D}_\phi$ if there exists a continuous linear operator $\phi'_\theta : \mathbb{D} \mapsto \mathbb{E}$ such that

$$\frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} \quad \to \quad \phi'_\theta(h), \tag{6.3}$$

as $n \to \infty$, for any scalar sequence $t_n \to 0$ and any $h, \{h_n\} \in \mathbb{D}$, with $h_n \to h$, and so that $\theta + t_n h_n \in \mathbb{D}_\phi$ for all $n$. It is left as an exercise to show that Hadamard differentiability is equivalent to *compact differentiability*, where compact differentiability satisfies

$$\sup_{h \in K, \theta + th \in \mathbb{D}_\phi} \left\| \frac{\phi(\theta + th) - \phi(\theta)}{t} - \phi'_\theta(h) \right\| \to 0, \quad \text{as } t \to 0, \tag{6.4}$$

for every compact $K \subset \mathbb{D}$. Hadamard differentiability can be refined by restricting the $h$ values to be in a set $\mathbb{D}_0 \subset \mathbb{D}$. More precisely, if in (6.3) it is required that $h_n \to h$ only for $h \in \mathbb{D}_0 \subset \mathbb{D}$, we say $\phi$ is *Hadamard-differentiable tangentially* to the set $\mathbb{D}_0$. There appears to be no easy way to refine compact differentiability in an equivalent manner.

A map $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ is *Fréchet-differentiable* if there exists a continuous linear operator $\phi'_\theta : \mathbb{D} \mapsto \mathbb{E}$ so that (6.4) holds uniformly in $h$ on bounded subsets of $\mathbb{D}$. This is equivalent to $\|\phi(\theta + h) - \phi(\theta) - \phi'_\theta(h)\| = o(\|h\|)$, as $\|h\| \to 0$. Since compact sets are bounded, Fréchet differentiability implies Hadamard differentiability. Fréchet differentiability will be needed for Z-estimator theory, while Hadamard differentiability is useful in the delta method. The following chain rule for Hadamard differentiability will also prove useful:

LEMMA 6.19 **(Chain rule)** *Assume $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}_\psi \subset \mathbb{E}$ is Hadamard differentiable at $\theta \in \mathbb{D}_\phi$ tangentially to $\mathbb{D}_0 \subset \mathbb{D}$, and $\psi : \mathbb{E}_\psi \subset \mathbb{E} \mapsto \mathbb{F}$ 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.** First, $\psi \circ \phi(\theta + th_t) - \psi \circ \phi(\theta) = \psi(\phi(\theta) + tk_t) - \psi(\phi(\theta))$, where $k_t = [\phi(\theta + th_t) - \phi(\theta)]/t$. Note that if $h \in \mathbb{D}_0$, then $k_t \to k \equiv \phi'_\theta(h) \in \phi'_\theta(\mathbb{D}_0)$, as $t \to 0$, by the Hadamard differentiability of $\phi$. Now, $[\psi(\phi(\theta) + tk_t) - \psi(\phi(\theta))]/t \to \psi'_{\phi(\theta)}(k)$ by the Hadamard differentiability of $\psi$.□

## 6.4 Proofs

**Proof of Theorem 6.2.** First assume $\overline{K} \subset \ell^\infty(T)$ is compact. Let $\rho(s,t) = \sup_{x \in K} |x(s) - x(t)|$. We will now establish that $(T, \rho)$ is totally bounded.

Fix $\eta > 0$ and cover $K$ with finitely many open balls of radius $\eta$, centered at $x_1, \ldots, x_k$. Partition $\mathbb{R}^k$ into cubes with edges of length $\eta$. For every such cube for which it is possible, choose at most one $t \in T$ so that $(x_1(t), \ldots, x_k(t))$ is in the cube. This results in a finite set $T_\eta \equiv \{t_1, \ldots, t_m\}$ in $T$ because $x_1, \ldots, x_k$ are uniformly bounded. For each $s \in T_\eta$, we have for all values of $t \in T$ for which $(x_1(t), \ldots, x_k(t))$ and $(x_1(s), \ldots, x_k(s))$ are in the same cube, that

$$\begin{aligned} \rho(s,t) &= \sup_{x \in K} |x(s) - x(t)| \\ &\leq \sup_{1 \leq j \leq k} |x_j(s) - x_j(t)| + 2 \sup_{x \in K} \inf_{1 \leq j \leq k} \sup_{u \in T} |x_j(u) - x(u)| \\ &< 3\eta. \end{aligned}$$

Thus the balls $\{t : \rho(t, t_1) < 3\eta\}, \ldots, \{t : \rho(t, t_m) < 3\eta\}$ completely cover $T$. Hence $(T, \rho)$ is totally bounded since $\eta$ was arbitrary. Also, by construction, the condition in Part (a.ii) of the theorem is satisfied. Combining this with the total boundedness of $(T, \rho)$ yields Condition (a.i). We have now obtained the fairly strong result that compactness of the closure of $K \subset \ell^\infty(T)$ implies that there exists a semimetric $\rho$ which makes $(T, \rho)$ totally bounded and which enables Conditions (a.i) and (a.ii) to be satisfied for $K$.

Now assume that the closure of $K \subset \ell^\infty(T)$ is $\sigma$-compact. This implies that there exists a sequence of compact sets $K_1 \subset K_2 \subset \cdots$ for which $\overline{K} = \cup_{i \geq 1} K_i$. The previous result yields for each $i \geq 1$, that there exists a semimetric $\rho_i$ making $T$ totally bounded and for which Conditions (a.i) and (a.ii) are satisfied for each $K_i$. Now let $\rho(s,t) = \sum_{i=1}^\infty 2^{-i} (\rho_i(s,t) \wedge 1)$. Fix $\eta > 0$, and select a finite integer $m$ so that $2^{-m} < \eta$. Cover $T$ with finitely many open $\rho_m$ balls of radius $\eta$, and let $T_\eta = \{t_1, \ldots, t_k\}$ be their centers. Because $\rho_1 \leq \rho_2 \leq \cdots$, there is for every $t \in T$ an $s \in T_\eta$ with $\rho(s,t) \leq \sum_{i=1}^m 2^{-i} \rho_i(s,t) + 2^{-m} \leq 2\eta$. Thus $(T, \rho)$ is totally bounded by $\rho$ since $\eta$ was arbitrary. Now, for any $x \in K$, $x \in K_m$ for some finite $m \geq 1$, and thus $x$ is both bounded and uniformly $\rho$-continuous since $\rho_m \leq 2^m \rho$. Hence $\sigma$-compactness of $K \subset \ell^\infty(T)$ implies $K \subset UC(T, \rho)$ for some semimetric $\rho$ making $T$ totally bounded. In the discussion preceding the statement of Theorem 6.2, we argued that $UC(T, \rho)$ is $\sigma$-compact whenever $(T, \rho)$ is totally bounded. Hence we have proven Part (b).

The only part of the proof which remains is to show that if $K \subset UC(T, \rho)$ satisfies Conditions (a.i) and (a.ii), then the closure of $K$ is compact. Assume Conditions (a.i) and (a.ii) hold for $K$. Define

$$m_\delta(x) \equiv \sup_{s,t \in T : \rho(s,t) < \delta} |x(s) - x(t)|$$

and $m_\delta \equiv \sup_{x \in K} m_\delta(x)$, and note that $m_\delta(x)$ is continuous in $x$ and that $m_{1/n}(x)$ is nonincreasing in $n$, with $\lim_{n \to \infty} m_{1/n}(x) = 0$. Choose $k < \infty$ large enough so that $m_{1/k} < \infty$. For every $\delta > 0$, let $T_\delta \subset T$ be a finite mesh

satisfying $\sup_{t\in T}\inf_{s\in T_\delta}\rho(s,t)<\delta$, and let $N_\delta$ be the number of points in $T_\delta$. Now, for any $t\in T$, $|x(t)|\le|x(t_0)|+|x(t)-x(t_0)|\le|x(t_0)|+N_{1/k}m_{1/k}$, and thus $\alpha\equiv\sup_{x\in K}\sup_{t\in T}|x(t)|<\infty$. For each $\epsilon>0$, pick a $\delta>0$ so that $m_\delta<\epsilon$ and an integer $n<\infty$ so that $\alpha/n\le\epsilon$. Let $U\equiv T_{1/(2\delta)}$ and define a "bracket" to be a finite collection $h=\{h_t:t\in U\}$ so that $h_t=-\alpha+j\alpha/n$, for some integer $1\le j\le 2n-1$, for each $t\in U$. Say that $x\in\ell^\infty(T)$ is "in" the bracket $h$, denoted $x\in h$, if $x(t)\in[h_t,h_t+\alpha/n]$ for all $t\in U$. Let $B(K)$ be the set of all brackets $h$ for which $x\in h$ for some $x\in K$. For each $h\in B(K)$, choose one and only one $x\in K$ with $x\in h$, discard duplicates, and denote the resulting set $X(K)$. It is not hard to verify that $\sup_{x\in K}\inf_{y\in X(K)}\|x-y\|_T<2\epsilon$, and thus the union of $2\epsilon$-balls with centers in $X(K)$ is a finite cover of $K$. Since $\epsilon$ is arbitrary, $K$ is totally bounded.□

**Proof of Lemma 6.3.** Begin by selecting a measurable sequence $U_m\ge T$ such that $\mathrm{E}\arctan U_m\downarrow\mathrm{E}^*\arctan T$, and set $T^*(\omega)=\lim_{m\to\infty}\inf_{1\le k\le m}U_k(\omega)$. This gives a measurable function $T^*$ taking values in $\bar{\mathbb{R}}$, with $T^*\ge T$, and $\mathrm{E}\arctan T^*=\mathrm{E}^*\arctan T$ by monotone convergence. For any measurable $U\ge T$, $\arctan U\wedge T^*\ge\arctan T$, and thus $\mathrm{E}\arctan U\wedge T^*\ge\mathrm{E}^*\arctan T=\mathrm{E}\arctan T^*$. However, $U\wedge T^*$ is trivially smaller than $T^*$, and since both quantities therefore have the same expectation, $\arctan U\wedge T^*=\arctan T^*$ a.s. Hence $T^*\le U$ a.s., and (i) and (ii) follow. When $\mathrm{E}T^*$ exists, it is larger than $\mathrm{E}^*T$ by (i) yet smaller by (ii), and thus $\mathrm{E}T^*=\mathrm{E}^*T$. When $\mathrm{E}^*T<\infty$, there exists a measurable $U\ge T$ with $\mathrm{E}U^+<\infty$, where $z^+$ is the positive part of $z$. Hence $\mathrm{E}(T^*)^+\le\mathrm{E}U^+$ and $\mathrm{E}T^*$ exists.□

**Proof of Lemma 6.6.** The second inequality in (i) is obvious. If $S$ and $U\ge S+T$ are both measurable, then $U-S\ge T$ and $U-S\ge T^*$ since $U-S$ is also measurable. Hence $U=(S+T)^*\ge S+T^*$ and the second inequality is an equality. Now $(S+T)^*\ge(S_*+T)^*=S_*+T^*$, and we obtain the first inequality. If $S$ is measurable, then $S_*=S^*$ and thus $S_*+T^*=S^*+T^*$. Part (ii) is left as an exercise. Part (iii) follows from the second inequality in (i) after relabeling and rearranging. Part (iv) follows from $S^*-T^*\le(S-T)^*\le|S-T|^*$ and then exchanging the roles of $S$ and $T$. For Part (v), it is clear that $(1\{T>c\})^*\ge1\{T^*>c\}$. If $U\ge1\{T>c\}$ is measurable, then $S=T^*1\{U\ge1\}+(T^*\wedge c)1\{U<1\}\ge T$ and is measurable. Hence $S\ge T^*$, and thus $T^*\le c$ whenever $U<1$. This trivially implies $1\{T^*>c\}=0$ when $U<1$, and thus $1\{T^*>c\}\le U$. Part (vi) is left as an exercise.

For Part (vii), $(S\vee T)^*\le S^*\vee T^*$ trivially. Let $U=(S\vee T)^*$ and note that $U$ is measurable and both $U\ge T$ and $U\ge S$. Hence both $U\ge T^*$ and $U\ge S^*$, and thus $(S\vee T)^*\ge S^*\vee T^*$, yielding the desired inequality. The inequality in (viii) is obvious. Assume $S$ is measurable and let $U=(S\wedge T)^*$. Clearly, $U\le S^*\wedge T^*$. Define $\tilde{T}\equiv U1\{U<S\}+T^*1\{U\ge S\}$; and note that $\tilde{T}\ge T^*$, since if $U<S$, then $T<S$ and thus $U\ge T$. Fix $\omega\in\Omega$. If $U<S$, then $S\wedge\tilde{T}=U$. If $U\ge S$, then $U=S$ since $U\le S$, and, furthermore, we will now show that $T^*\ge S$. If it were not true, then $T^*<S$ and

$U \le S \wedge T^* < S$, which is clearly a contradiction. Thus when $U \ge S$, $U = S = S \wedge \tilde{T} \ge S \wedge T^*$. Hence $U = S \wedge \tilde{T} \ge S \wedge T^*$ a.s., and the desired equality in (viii) follows.□

**Proof of Lemma 6.7.** The first part of (i) is a consequence of the following chain of equalities: $1\{(A \cup B)^*\} = (1\{A \cup B\})^* = (1\{A\} \vee 1\{B\})^* = (1\{A\})^* \vee (1\{B\})^* = 1\{A^*\} \vee 1\{B^*\} = 1\{A^* \cup B^*\}$. The first and fourth equalities follow from (ii) of Lemma 6.5, the second and fifth equalities follow directly, and the third equality follows from (vii) of Lemma 6.6. For the second part of (i), we have $(A \cap B)_* = \Omega - [(\Omega - A) \cup (\Omega - B)]^* = \Omega - (\Omega - A)^* \cup (\Omega - B)^* = \Omega - (\Omega - A_*) \cup (\Omega - B_*) = A_* \cap B_*$, where the second equality follows from the first part of (i).

The inclusions in Part (ii) are obvious. Assume $A$ is measurable. Then (viii) of Lemma 6.6 can be used to validate the following string of equalities: $1\{(A \cap B)^*\} = (1\{A \cap B\})^* = (1\{A\} \wedge 1\{B\})^* = (1\{A\})^* \wedge (1\{B\})^* = 1\{A^*\} \wedge 1\{B^*\} = 1\{A^* \cap B^*\}$. Thus $(A \cap B)^* = A^* \cap B^*$. By symmetry, this works whether $A$ or $B$ is measurable. The proof that $(A \cup B)_* = A_* \cup B_*$ when either $A$ or $B$ is measurable is left as an exercise. The proof of Part (iii) is also left as an exercise.□

**Proof of Lemma 6.8.** All of the inequalities follow from the definitions. For the equality in A(i), assume that $\phi$ is nondecreasing and left-continuous. Define $\phi^{-1}(u) = \inf\{t : \phi(t) \ge u\}$, and note that $\phi(t) > u$ if and only if $t > \phi^{-1}(u)$. Thus, for any $c \in \mathbb{R}$, $1\{\phi(T^*) > c\} = 1\{T^* > \phi^{-1}(c)\} = \big(1\{T > \phi^{-1}(c)\}\big)^* = (1\{\phi(T) > c\})^* = 1\big\{[\phi(T)]^* > c\big\}$. The second and fourth equalities follow from (v) of Lemma 6.6. Hence $\phi(T^*) = [\phi(T)]^*$. For the equality in A(ii), assume that $\phi$ is nondecreasing and right-continuous; and define $\phi^{-1}(u) = \sup\{t : \phi(t) \le u\}$. Note that $\phi(t) \ge u$ if and only if $t \ge \phi^{-1}(u)$. The proof proceeds in the same manner as for A(i), only Part (vi) in Lemma 6.6 is used in place of Part (v). We leave the proof of Part B as an exercise.□

**Proof of Lemma 6.11.** Clearly, $\liminf T_n^* \le \limsup T_n^* \le T^*$. Conversely, $\liminf T_n^* \ge \liminf T_n = T$ and is measurable, and thus $\liminf T_n^* \ge T^*$. Hence $T_n^* \uparrow T^*$. Now $\mathrm{E}^* T_n^* = \mathrm{E} T_n^* \uparrow \mathrm{E} T^*$ by monotone convergence for measurable maps. Note we are allowing $+\infty$ as a possible value for $\mathrm{E}^* T_n$, for some $n$, or $\mathrm{E}^* T$.□

**Proof of Lemma 6.12.** Since $|T| \le |T_n| + |T - T_n|$ for all $n$, we have $|T - T_n|^* \le 2S^*$ a.s. Fix $\epsilon > 0$. Since $\mathrm{E}^* S < \infty$, there exists a $0 < k < \infty$ so that $\mathrm{E}[S^* 1\{S^* > k\}] \le \epsilon/2$. Thus $\mathrm{E}^* |T - T_n| \le \mathrm{E} k \wedge |T - T_n|^* + 2\mathrm{E}[S^* 1\{S^* > k\}] \to \epsilon$. The result now follows since $\epsilon$ was arbitrary.□

## 6.5 Exercises

**6.5.1.** Show that Part (iii) in the definition of $\sigma$-field can be replaced, without really changing the definition, by the following: The countable intersection $\bigcap_{j=1}^{\infty} U_j \in \mathcal{A}$ whenever $U_j \in \mathcal{A}$ for all $j \geq 1$.

**6.5.2.** Show the following:

(a) For a metric space $\mathbb{D}$ and set $A \in \mathbb{D}$, the closure $\overline{A}$ consists of all limits of sequences $\{x_n\} \in A$.

(b) Two metrics $d_1$ and $d_2$ on a set $\mathbb{D}$ are equivalent if and only if we have the following for any sequence $\{x_j\} \in \mathbb{D}$, as $n, m \to \infty$: $d_1(x_n, x_m) \to 0$ if and only if $d_2(x_n, x_m) \to 0$.

(c) A function $f : \mathbb{D} \mapsto \mathbb{E}$ between two metric spaces is continuous (in the topological sense) if and only if, for all $x \in \mathbb{D}$ and all sequences $\{x_n\} \in \mathbb{D}$, $f(x_n) \to f(x)$ whenever $x_n \to x$.

**6.5.3.** Verify the implication (ii) $\Rightarrow$ (i) in Lemma 6.1.

**6.5.4.** Show that if a subset $K$ of a metric space is not totally bounded, then it is possible to construct a sequence $\{x_n\} \in K$ which has no Cauchy subsequences.

**6.5.5.** Show that the following spaces are complete with respect to the uniform metric:

(a) $C[a, b]$ and $D[a, b]$;

(b) $UC(T, \rho)$ and $C(\overline{T}, \rho)$, where $(T, \rho)$ is a totally bounded semimetric space;

(c) $UC(T, \rho)$ and $\ell^{\infty}(T)$, where $T$ is an arbitrary set.

**6.5.6.** Show that a uniformly continuous function $f : T \mapsto \mathbb{R}$, where $T$ is totally bounded, has a unique continuous extension to $\overline{T}$.

**6.5.7.** Let $C_L[0, 1] \subset C[0, 1]$ be the space of all Lipschitz-continuous functions on $[0, 1]$, and endow it with the uniform metric:

(a) Show that $C_L[0, 1]$ is dense in $C[0, 1]$.

(b) Show that no open ball in $C[0, 1]$ is contained in $C_L[0, 1]$.

(c) Show that $C_L[0, 1]$ is not complete.

(d) Show that for $0 < c < \infty$, the set

$$\{f \in C_L[0, 1] : |f(x)| \leq c \text{ and } |f(x) - f(y)| \leq c|x - y|, \ \forall x, y \in [0, 1]\}$$

is compact.

**6.5.8.** A collection $\mathcal{F}$ of maps $f : \mathbb{D} \mapsto \mathbb{E}$ between metric spaces, with respective Borel $\sigma$-fields $\mathcal{D}$ and $\mathcal{E}$, can generate a (possibly) new $\sigma$-field for $\mathbb{D}$ by considering all inverse images $f^{-1}(A)$, for $f \in \mathcal{F}$ and $A \in \mathcal{E}$. Show that the $\sigma$-field $\sigma_p$ generated by the coordinate projections $x \mapsto x(t)$ on $C[a, b]$ is equal to the Borel $\sigma$-field $\sigma_c$ generated by the uniform norm. Hint: Show first that continuity of the projection maps implies $\sigma_p \subset \sigma_c$. Second, show that the open balls in $\sigma_c$ can be created from countable set operations on sets in $\sigma_p$.

**6.5.9.** Show that the following metrics generate the product topology on $\mathbb{D} \times \mathbb{E}$, where $d$ and $e$ are the respective metrics for $\mathbb{D}$ and $\mathbb{E}$:

(i) $\rho_1((x_1, y_1), (x_2, y_2)) \equiv d(x_1, x_2) + e(y_1, y_2)$.

(ii) $\rho_2((x_1, y_1), (x_2, y_2)) \equiv \sqrt{d^2(x_1, x_2) + e^2(y_1, y_2)}$.

**6.5.10.** For any map $T : \Omega \mapsto \bar{\mathbb{R}}$, show that $\mathrm{E}_* T$ is the supremum of all $\mathrm{E}U$, where $U \leq T$, $U : \Omega \mapsto \bar{\mathbb{R}}$ measurable and $\mathrm{E}U$ exists. Show also that for any set $B \in \Omega$, $P_*(B)$ is the supremum of all $P(A)$, where $A \subset B$ and $A \in \mathcal{A}$.

**6.5.11.** Use Lemma 6.3 to prove Lemma 6.4.

**6.5.12.** Prove Parts (ii) and (vi) of Lemma 6.6 using Parts (i) and (v), respectively.

**6.5.13.** Let $S, T : \Omega \mapsto \bar{\mathbb{R}}$ be arbitrary. Show the following:

(a) $|S^* - T_*| \leq |S - T|^* + (S^* - S_*) \wedge (T^* - T_*)$;

(b) $|S - T|^* \leq (S^* - T_*) \vee (T^* - S_*) \leq |S - T|^* + (S^* - S_*) \wedge (T^* - T_*)$.

**6.5.14.** Finish the proof of Lemma 6.7:

(a) Show that $(A \cup B)_* = A_* \cup B_*$ if either $A$ or $B$ is measurable.

(b) Prove Part (iii) of the lemma.

**6.5.15.** Prove Part B of Lemma 6.8 using the approach outlined in the proof of Part A.

**6.5.16.** Prove the following "converse" to Jensen's inequality:

LEMMA 6.20 **(converse to Jensen's inequality)** *Let $T : \Omega \mapsto \mathbb{R}$ be an arbitrary map, with $\mathrm{E}^*|T| < \infty$, and assume $\phi : \mathbb{R} \mapsto \mathbb{R}$ is concave. Then*

*(a)* $\mathrm{E}_* \phi(T) \leq \phi(\mathrm{E}_* T) \wedge \phi(\mathrm{E}^* T)$*;*

*(b) if $\phi$ is also monotone,* $\mathrm{E}^* \phi(T) \leq \phi(\mathrm{E}_* T) \vee \phi(\mathrm{E}^* T)$.

**6.5.17.** In the proof of Proposition 6.15, show that (i) implies (ii).

**6.5.18.** Show that Hadamard and compact differentiability are equivalent.

## 6.6 Notes

Much of the material in Section 6.2 is an amalgamation of several concepts and presentation styles found in Chapter 2 of Billingsley (1986), Sections 1.3 and 1.7 of van der Vaart and Wellner (1996), Section 18.1 of van der Vaart (1998), and in Chapter 1 of both Rudin (1987) and Rudin (1991). A nice proof of the equivalence of the several definitions of compactness can by found in Appendix I of Billingsley (1968).

Many of the ideas in Section 6.3 come from Chapter 1.2 of van der Vaart and Wellner (1996), abbreviated VW hereafter. Lemma 6.3 corresponds to Lemma 1.2.1 of VW, Lemma 6.4 is given in Exercise 1.2.1 of VW, and Parts (i), (ii) and (iii) correspond to Lemma 1.2.3 of VW. Most of Lemma 6.6 is given in Lemma 1.2.2 of VW, although the first inequalities in Parts (i) and (ii), as well as Part (vi), are new. Lemma 6.7 is given in Exercise 1.2.15 in VW. Lemmas 6.11 and 6.12 correspond to Exercises 1.2.3 and 1.2.4 of VW, respectively, after some modification. Also, Lemmas 6.13 and 6.14 correspond to Lemmas 1.2.5 and 1.2.6 of VW, respectively.

Much of the material on linear operators can be found in Appendix A.1 of Bickel, Klaassen, Ritov and Wellner (1998), and in Chapter 2 of Kress (1999). Lemma 6.16 is Proposition 7, Parts A and B, in Appendix A.1 of Bickel, et al. (1998). The presentation on functional differentiation is motivated by the first few pages in Chapter 3.9 of VW, and the chain rule (Lemma 6.19) is Lemma 3.9.3 of VW.

# 7
# Stochastic Convergence

In this chapter, we study concepts and theory useful in understanding the limiting behavior of stochastic processes. We begin with a general discussion of stochastic processes in metric spaces. The focus of this discussion is on measurable stochastic processes since most limits of empirical processes in statistical applications are measurable. We next discuss weak convergence both in general and in the specific case of bounded stochastic processes. One of the interesting aspects of the approach we take to weak convergence is that the processes studied need not be measurable except in the limit. This is useful in applications since many empirical processes in statistics are not measurable with respect to the uniform metric. The final section of this chapter considers other modes of convergence, such as in probability and outer almost surely, and their relationships to weak convergence.

## 7.1 Stochastic Processes in Metric Spaces

In this section, we introduce several important concepts about stochastic processes in metric spaces. Recall that for a stochastic process $\{X(t), t \in T\}$, $X(t)$ is a measurable real random variable for each $t \in T$ on a probability space $(\Omega, \mathcal{A}, P)$. The sample paths of such a process typically reside in the metric space $\mathbb{D} = \ell^\infty(T)$ with the uniform metric. Often, however, when $X$ is viewed as a map from $\Omega$ to $\mathbb{D}$, it is no longer Borel measurable. A classic example of this duality comes from Billingsley (1968, Pages 152–153). The example hinges on the fact that there exists a set $H \subset [0, 1]$ which is not a

Borel set. Define the stochastic process $X(t) = 1\{U \leq t\}$, where $t \in [0,1]$ and $U$ is uniformly distributed on $[0,1]$. The probability space is $(\Omega, \mathcal{B}, P)$, where $\Omega = [0,1]$, $\mathcal{B}$ are the Borel sets on $[0,1]$, and $P$ is the uniform probability measure on $[0,1]$. A natural metric space for the sample paths of $X$ is $\ell^\infty([0,1])$. Define the set $A = \cup_{s \in H} B_s(1/2)$, where $B_s(1/2)$ is the uniform open ball of radius $1/2$ around the function $t \mapsto f_s(t) \equiv 1\{t \leq s\}$. Since $A$ is an open set in $\ell^\infty([0,1])$, and since the uniform distance between $f_{s_1}$ and $f_{s_2}$ is 1 whenever $s_1 \neq s_2$, the set $\{\omega \in \Omega : X(\omega) \in A\}$ equals $H$. Since $H$ is not a Borel set, $X$ is not Borel measurable.

This lack of measurability is actually the usual state for most of the empirical processes we are interested in studying, especially since most of the time the uniform metric is the natural choice of metric. Many of the associated technical difficulties can be resolved through the use of outer measure and outer expectation as defined in the previous chapter and which we will utilize in our study of weak convergence. In contrast, most of the limiting processes we will be studying are, in fact, Borel measurable. For this reason, a brief study of Borel measurable processes is valuable. The following lemma, for example, provides two ways of establishing equivalence between Borel probability measures. Recall from Section 2.2.3 that $BL_1(\mathbb{D})$ is the set of all functions $f : \mathbb{D} \mapsto \mathbb{R}$ bounded by 1 and with Lipschitz norm bounded by 1, i.e., with $|f(x) - f(y)| \leq d(x,y)$ for all $x, y \in \mathbb{D}$. When the choice of metric space is clear by the context, we will sometimes use the abbreviated notation $BL_1$ as was done in Chapter 2 . Define also a *vector lattice* $\mathcal{F} \subset C_b(\mathbb{D})$ to be a vector space for which if $f \in \mathcal{F}$ then $f \vee 0 \in \mathcal{F}$. We also say that a set $\mathcal{F}$ of real functions on $\mathbb{D}$ *separates points* of $\mathbb{D}$ if, for any $x, y \in \mathbb{D}$ with $x \neq y$, there exists $f \in \mathcal{F}$ such that $f(x) \neq f(y)$. We are now ready for the lemma:

LEMMA 7.1 *Let $L_1$ and $L_2$ be Borel probability measures on a metric space $\mathbb{D}$. The following are equivalent:*

*(i)* $L_1 = L_2$.

*(ii)* $\int f dL_1 = \int f dL_2$ *for all* $f \in C_b(\mathbb{D})$.

*If $L_1$ and $L_2$ are also separable, then (i) and (ii) are both equivalent to*

*(iii)* $\int f dL_1 = \int f dL_2$ *for all* $f \in BL_1$.

*Moreover, if $L_1$ and $L_2$ are also tight, then (i)–(iii) are all equivalent to*

*(iv)* $\int f dL_1 = \int f dL_2$ *for all $f$ in a vector lattice $\mathcal{F} \subset C_b(\mathbb{D})$ that both contains the constant functions and separates points in $\mathbb{D}$.*

The proof is given in Section 7.4. We say that two Borel random maps $X$ and $X'$, with respective laws $L$ and $L'$, are *versions* of each other if $L = L'$.

In addition to being Borel measurable, most of the limiting stochastic processes of interest are *tight*. A Borel probability measure $L$ on a metric

space $\mathbb{D}$ is tight if for every $\epsilon > 0$, there exists a compact $K \subset \mathbb{D}$ so that $L(K) \geq 1 - \epsilon$. A Borel random map $X : \Omega \mapsto \mathbb{D}$ is tight if its law $L$ is tight. Tightness is equivalent to there being a $\sigma$-compact set that has probability 1 under $L$ or $X$. $L$ or $X$ is *separable* if there is a measurable and separable set which has probability 1. $L$ or $X$ is *Polish* if there is a measurable Polish set having probability 1. Note that tightness, separability and Polishness are all topological properties and do not depend on the metric. Since both $\sigma$-compact and Polish sets are also separable, separability is the weakest of the three properties. Whenever we say $X$ has any one of these three properties, we tacetly imply that $X$ is also Borel measurable.

On a complete metric space, tightness, separability and Polishness are equivalent. This equivalence for Polishness and separability follows from the definitions. To see the remaining equivalence, assume $L$ is separable. By completeness, there is a $\mathbb{D}_0 \subset \mathbb{D}$ having probability 1 which is both separable and closed. Fix any $\epsilon \in (0, 1)$. By separability, there exists a sequence $\{x_k\} \in \mathbb{D}_0$ which is dense in $\mathbb{D}_0$. For every $\delta > 0$, the union of the balls of radius $\delta$ centered on the $\{x_k\}$ covers $\mathbb{D}_0$. Hence for every integer $j \geq 1$, there exists a finite collection of balls of radius $1/j$ whose union $D_j$ has probability $\geq 1 - \epsilon/2^j$. Thus the closure of the intersection $\cap_{j \geq 1} D_j$ is totally bounded and has probability $\geq 1 - \epsilon$. Since $\epsilon$ is arbitrary, $L$ is tight.

For a stochastic process $\{X(t), t \in T\}$, where $(T, \rho)$ is a separable, semimetric space, there is another meaning for separable. $X$ is *separable* (as a stochastic process) if there exists a countable subset $S \subset T$ and a null set $N$ so that, for each $\omega \notin N$ and $t \in T$, there exists a sequence $\{s_m\} \in S$ with $\rho(s_m, t) \to 0$ and $|X(s_m, \omega) - X(t, \omega)| \to 0$. It turns out that many of the empirical processes we will be studying are separable in this sense, even though they are not Borel measurable and therefore cannot satisfy the other meaning for separable. Throughout the remainder of the book, the distinction between these two definitions will either be explicitly stated or made clear by the context.

Most limiting processes $X$ of interest will reside in $\ell^\infty(T)$, where the index set $T$ is often a class of real functions $\mathcal{F}$ with domain equal to the sample space. When such limiting processes are tight, the following lemma demands that $X$ resides on $UC(T, \rho)$, where $\rho$ is some semimetric making $T$ totally bounded, with probability 1:

LEMMA 7.2 *Let $X$ be a Borel measurable random element in $\ell^\infty(T)$. Then the following are equivalent:*

*(i) $X$ is tight.*

*(ii) There exists a semimetric $\rho$ making $T$ totally bounded and for which $X \in UC(T, \rho)$ with probability 1.*

*Furthermore, if (ii) holds for any $\rho$, then it also holds for the semimetric* $\rho_0(s, t) \equiv \mathrm{E} \arctan |X(s) - X(t)|$.

The proof is given in Section 7.4. A nice feature of tight processes in $\ell^\infty(T)$ is that the laws of such processes are completely defined by their finite-dimensional marginal distributions $(X(t_1), \ldots, X(t_k))$, where $t_1, \ldots, t_k \in T$ and $k \geq 1$ is an integer:

LEMMA 7.3 *Let $X$ and $Y$ be tight, Borel measurable stochastic processes in $\ell^\infty(T)$. Then the Borel laws of $X$ and $Y$ are equal if and only if all corresponding finite-dimensional marginal distributions are equal.*

**Proof.** Consider the collection $\mathcal{F} \subset C_b(\mathbb{D})$ of all functions $f : \ell^\infty(T) \mapsto \mathbb{R}$ of the form $f(x) = g(x(t_1), \ldots, x(t_k))$, where $g \in C_b(\mathbb{R}^k)$ and $k \geq 1$ is an integer. We leave it as an exercise to show that $\mathcal{F}$ is a vector lattice, an algebra, and separates points of $\ell^\infty(T)$. The desired result now follows from Lemma 7.1.□

While the semimetric $\rho_0$ defined in Lemma 7.2 is always applicable when $X$ is tight, it is frequently not the most convenient semimetric to work with. The family of semimetrics $\rho_p(s,t) \equiv (\mathrm{E}|X(s) - X(t)|^p)^{1/(p \vee 1)}$, for some choice of $p \in (0, \infty)$, is sometimes more useful. There is an interesting link between $\rho_p$ and other semimetrics for which Lemma 7.2 holds. For a process $X$ in $\ell^\infty(T)$ and a semimetric $\rho$ on $T$, we say that $X$ is *uniformly $\rho$-continuous in pth mean* if $\mathrm{E}|X(s_n) - X(t_n)|^p \to 0$ whenever $\rho(s_n, t_n) \to 0$. The following lemma is a conclusion from Lemma 1.5.9 of van der Vaart and Wellner (1996) (abbreviated VW hereafter), and we omit the proof:

LEMMA 7.4 *Let $X$ be a tight Borel measurable random element in $\ell^\infty(T)$, and let $\rho$ be any semimetric for which (ii) of Lemma 7.2 holds. If $X$ is $\rho$-continuous in pth mean for some $p \in (0, \infty)$, then (ii) of Lemma 7.2 also holds for the semimetric $\rho_p$.*

Perhaps the most frequently occurring limiting process in $\ell^\infty(T)$ is a *Gaussian* process. A stochastic process $\{X(t), t \in T\}$ is Gaussian if all finite-dimensional marginals $\{X(t_1), \ldots, X(t_k)\}$ are multivariate normal. If a Gaussian process $X$ is tight, then by Lemma 7.2, there is a semimetric $\rho$ making $T$ totally bounded and for which the sample paths $t \mapsto X(t)$ are uniformly $\rho$-continuous. An interesting feature of Gaussian processes is that this result implies that the map $t \mapsto X(t)$ is uniformly $\rho$-continuous in $p$th mean for all $p \in (0, \infty)$. To see this, take $p = 2$, and note that $|X(s_n) - X(t_n)| \to 0$ in probability if and only if $\mathrm{E}|X(s_n) - X(t_n)|^2 \to 0$. Thus whenever $\rho(s_n, t_n) \to 0$, $\mathrm{E}|X(s_n) - X(t_n)|^2 \to 0$ and hence also $\mathrm{E}|X(s_n) - X(t_n)|^p \to 0$ for any $p \in (0, \infty)$ since $X(s_n) - X(t_n)$ is normally distributed for all $n \geq 1$. Lemma 7.4 now implies that tightness of a Gaussian process is equivalent to $T$ being totally bounded by $\rho_p$ with almost all sample paths of $X$ being uniformly $\rho_p$-continuous for all $p \in (0, \infty)$.

For a general Banach space $\mathbb{D}$, a Borel measurable random element $X$ on $\mathbb{D}$ is Gaussian if and only if $f(X)$ is Gaussian for every continuous, linear map $f : \mathbb{D} \mapsto \mathbb{R}$. When $\mathbb{D} = \ell^\infty(T)$ for some set $T$, this definition appears to contradict the definition of Gaussianity given in the preceding

paragraph, since now we are using all continuous linear functionals instead of just linear combinations of coordinate projections. These two definitions are not really reconcilable in general, and so some care must be taken in reading the literature to determine the appropriate context. However, when the process in question is tight, the two definitions are equivalent, as verified in the following proposition:

PROPOSITION 7.5 *Let $X$ be a tight, Borel measurable map into $\ell^\infty(T)$. Then the following are equivalent:*

*(i) The vector $(X_{t_1}, \ldots, X_{t_k})$ is multivariate normal for every finite set $\{t_1, \ldots, t_k\} \subset T$.*

*(ii) $\phi(X)$ is Gaussian for every continuous, linear map $\phi : \ell^\infty(T) \mapsto \mathbb{R}$.*

*(iii) $\phi(X)$ is Gaussian for every continuous, linear map $\phi : \ell^\infty(T) \mapsto \mathbb{D}$ into any Banach space $\mathbb{D}$.*

**Proof.** The proof that (i)⇒(ii) is given in the proof of Lemma 3.9.8 of VW, and we omit the details here. Now assume (ii), and fix any Banach space $\mathbb{D}$ and any continuous, linear map $\phi : \ell^\infty(T) \mapsto \mathbb{D}$. Now for any continuous, linear map $\psi : \mathbb{D} \mapsto \mathbb{R}$, the composition map $\psi \circ \phi : \ell^\infty(T) \mapsto \mathbb{R}$ is continuous and linear, and thus by (ii) we have that $\psi(\phi(X))$ is Gaussian. Since $\psi$ is arbitrary, we have by the definition of a Gaussian process on a Banach space that $\phi(X)$ is Gaussian. Since both $\mathbb{D}$ and $\phi$ were also arbitrary, conclusion (iii) follows. Finally, (iii)⇒(i) since multivariate coordinate projections are special examples of continuous, linear maps into Banach spaces.□

## 7.2 Weak Convergence

We first discuss the general theory of weak convergence in metric spaces and then discuss results for the special metric space of uniformly bounded functions, $\ell^\infty(T)$, for an arbitrary index set $T$. This last space is where most—if not all—of the action occurs for statistical applications of empirical processes.

### *7.2.1 General Theory*

The extremely important concept of weak convergence of sequences arises in many areas of statistics. To be as flexible as possible, we allow the probability spaces associated with the sequences to change with $n$. Let $(\Omega_n, \mathcal{A}_n, P_n)$ be a sequence of probability spaces and $X_n : \Omega_n \mapsto \mathbb{D}$ a sequence of maps. We say that $X_n$ *converges weakly* to a Borel measurable $X : \Omega \mapsto \mathbb{D}$ if

$$\mathrm{E}^* f(X_n) \to \mathrm{E} f(X), \quad \text{for every } f \in C_b(\mathbb{D}). \tag{7.1}$$

If $L$ is the law of $X$, (7.1) can be reexpressed as

$$\mathrm{E}^* f(X_n) \to \int_\Omega f(x) dL(x), \quad \text{for every } f \in C_b(\mathbb{D}).$$

This weak convergence is denoted $X_n \rightsquigarrow X$ or, equivalently, $X_n \rightsquigarrow L$. Weak convergence is equivalent to "convergence in distribution" and "convergence in law." By Lemma 7.1, this definition of weak convergence ensures that the limiting distributions are unique. Note that the choice of probability spaces $(\Omega_n, \mathcal{A}_n, P_n)$ is important since these dictate the outer expectation used in the definition of weak convergence. In most of the settings discussed in this book, $\Omega_n = \Omega$ for all $n \geq 1$. Some important exceptions to this rule will be discussed in Chapter 11. Fortunately, even in those settings where $\Omega_n$ does change with $n$, one can frequently readjust the probability spaces so that the sample spaces are all the same. It is also possible to generalize the concept of weak convergence of sequences to weak convergence of nets as done in VW, but we will restrict ourselves to sequences throughout this book.

The forgoing definition of weak convergence does not obviously appear to be related to convergence of probabilities, but this is in fact true for the probabilities of sets $B \subset \Omega$ which have boundaries $\delta B$ satisfying $L(\delta B) = 0$. Here and elsewhere, we define the boundary $\delta B$ of a set $B$ in a topological space to be the closure of $B$ minus the interior of $B$. Several interesting equivalent formulations of weak convergence on a metric space $\mathbb{D}$ are given in the following portmanteau theorem:

THEOREM 7.6 *(Portmanteau) The following are equivalent:*

*(i)* $X_n \rightsquigarrow L$;

*(ii)* $\liminf \mathrm{P}_*(X_n \in G) \geq L(G)$ *for every open* $G$;

*(iii)* $\limsup \mathrm{P}^*(X_n \in F) \leq L(F)$ *for every closed* $F$;

*(iv)* $\liminf \mathrm{E}_* f(X_n) \geq \int_\Omega f(x) dL(x)$ *for every lower semicontinuous* $f$ *bounded below;*

*(v)* $\limsup \mathrm{E}^* f(X_n) \leq \int_\Omega f(x) dL(x)$ *for every upper semicontinuous* $f$ *bounded above;*

*(vi)* $\lim \mathrm{P}^*(X_n \in B) = \lim \mathrm{P}_*(X_n \in B) = L(B)$ *for every Borel* $B$ *with* $L(\delta B) = 0$;

*(vii)* $\liminf \mathrm{E}_* f(X_n) \geq \int_\Omega f(x) dL(x)$ *for every bounded, Lipschitz continuous, nonnegative* $f$.

*Furthermore, if* $L$ *is separable, then (i)–(vii) are also equivalent to*

*(viii)* $\sup_{f \in BL_1} |\mathrm{E}^* f(X_n) - \mathrm{E} f(X)| \to 0$.

The proof is given in Section 7.4. Depending on the setting, one or more of these alternative definitions will prove more useful than the others. For example, definition (vi) is probably the most intuitive from a statistical point of view, while definition (viii) is convenient for studying certain properties of the bootstrap.

Another very useful result is the continuous mapping theorem:

THEOREM 7.7 *(Continuous mapping) Let $g : \mathbb{D} \mapsto \mathbb{E}$ be continuous at all points in $\mathbb{D}_0 \subset \mathbb{D}$, where $\mathbb{D}$ and $\mathbb{E}$ are metric spaces. Then if $X_n \rightsquigarrow X$ in $\mathbb{D}$, with $\mathrm{P}_*(X \in \mathbb{D}_0) = 1$, then $g(X_n) \rightsquigarrow g(X)$.*

The proof is given in Section 7.4. As mentioned in Chapter 2, a common application of this theorem is in the construction of confidence bands based on the supremum distance.

A potential issue is that there may sometimes be more than one choice of metric space $\mathbb{D}$ to work with in a given weak convergence setting. For example, if we are studying weak convergence of the usual empirical process $\sqrt{n}(\hat{F}_n(t) - F(t))$ based on data in $[0, 1]$, we could let $\mathbb{D}$ be either $\ell^\infty([0, 1])$ or $D[0, 1]$. The following lemma tells us that the choice of metric space is generally not a problem. Recall from Chapter 6 that for a topological space $(X, \mathcal{O})$, the relative topology on $A \subset X$ consists of the open sets $\{A \cap B : B \in \mathcal{O}\}$.

LEMMA 7.8 *Let the metric spaces $\mathbb{D}_0 \subset \mathbb{D}$ have the same metric, and assume $X$ and $X_n$ reside in $\mathbb{D}_0$. Then $X_n \rightsquigarrow X$ in $\mathbb{D}_0$ if and only if $X_n \rightsquigarrow X$ in $\mathbb{D}$.*

**Proof.** Since any set $B_0 \in \mathbb{D}_0$ is open if and only if it is of the form $B \cap \mathbb{D}_0$ for some open $B$ in $\mathbb{D}$, the result follows from Part (ii) of the portmanteau theorem.□

Recall from Chapter 2 that a sequence $X_n$ is asymptotically measurable if and only if

$$\mathrm{E}^* f(X_n) - \mathrm{E}_* f(X_n) \quad \to \quad 0, \tag{7.2}$$

for all $f \in C_b(\mathbb{D})$. An important, related concept is that of *asymptotic tightness*. A sequence $X_n$ is asymptotically tight if for every $\epsilon > 0$, there is a compact $K$ so that $\liminf \mathrm{P}_*(X_n \in K^\delta) \geq 1 - \epsilon$, for every $\delta > 0$, where for a set $A \subset \mathbb{D}$, $A^\delta = \{x \in \mathbb{D} : d(x, A) < \delta\}$ is the "$\delta$-enlargement" around $A$. The following lemma tells us that when $X_n$ is asymptotically tight, we can determine asymptotic measurability by verifying (7.2) only for a subset of functions in $C_b(\mathbb{D})$. For the purposes of this lemma, an *algebra* $\mathcal{F} \subset C_b(\mathbb{D})$ is a vector space for which if $f, g \in \mathcal{F}$ then $fg \in \mathcal{F}$.

LEMMA 7.9 *Assume the sequence $X_n$ is asymptotically tight and that (7.2) holds for all $f$ in a subalgebra $\mathcal{F} \subset C_b(\mathbb{D})$ that separates points of $\mathbb{D}$. Then $X_n$ is asymptotically measurable.*

We omit the proof of this lemma, but it can be found in Chapter 1.3 of VW.

When $\mathbb{D}$ is a Polish space and $X_n$ and $X$ are both Borel measurable, tightness of $X_n$ for each $n \geq 1$ plus asymptotic tightness is equivalent to the concept of *uniform tightness* used in the classical theory of weak convergence (see p. 37, Billingsley, 1968). More precisely, a Borel measurable sequence $\{X_n\}$ is uniformly tight if for every $\epsilon > 0$, there is a compact $K$ so that $\mathrm{P}(X_n \in K) \geq 1 - \epsilon$ for all $n \geq 1$. The following is a more formal statement of the equivalence we are describing:

LEMMA 7.10 *Assume $\mathbb{D}$ is a Polish space and that the maps $X_n$ and $X$ are Borel measurable. Then $\{X_n\}$ is uniformly tight if and only if $X_n$ is tight for each $n \geq 1$ and $\{X_n\}$ is asymptotically tight.*

The proof is given in Section 7.4. Because $X_n$ will typically not be measurable in many of the applications of interest to us, uniform tightness will not prove as useful a concept as asymptotic tightness.

Two good properties of asymptotic tightness are that it does not depend on the metric chosen—only on the topology—and that weak convergence often implies asymptotic tightness. The first of these two properties are verified in the following lemma:

LEMMA 7.11 *$X_n$ is asymptotically tight if and only if for every $\epsilon > 0$ there exists a compact $K$ so that $\liminf \mathrm{P}_*(X_n \in G) \geq 1 - \epsilon$ for every open $G \supset K$.*

**Proof.** Assume first that $X_n$ is asymptotically tight. Fix $\epsilon > 0$, and let the compact set $K$ satisfy $\liminf \mathrm{P}_*(X_n \in K^\delta) \geq 1 - \epsilon$ for every $\delta > 0$. If $G \supset K$ is open, then there exists a $\delta_0 > 0$ so that $G \supset K^{\delta_0}$. If this were not true, then there would exist a sequence $\{x_n\} \notin G$ so that $d(x_n, K) \to 0$. This implies the existence of a sequence $\{y_n\} \in K$ so that $d(x_n, y_n) \to 0$. Thus, since $K$ is compact and the complement of $G$ is closed, there is a subsequence $n'$ and a point $y \notin G$ so that $d(y_{n'}, y) \to 0$, but this is impossible. Hence $\liminf \mathrm{P}_*(X_n \in G) \geq 1 - \epsilon$. Now assume that $X_n$ satisfies the alternative definition. Fix $\epsilon > 0$, and let the compact set $K$ satisfy $\liminf \mathrm{P}_*(X_n \in G) \geq 1 - \epsilon$ for every open $G \supset K$. For every $\delta > 0$, $K^\delta$ is an open set. Thus $\liminf \mathrm{P}_*(X_n \in K^\delta) \geq 1 - \epsilon$ for every $\delta > 0$.□

The second good property of asymptotic tightness is given in the second part of the following lemma, the first part of which gives the necessity of asymptotic measurability for weakly convergent sequences:

LEMMA 7.12 *Assume $X_n \rightsquigarrow X$. Then*

*(i) $X_n$ is asymptotically measurable.*

*(ii) $X_n$ is asymptotically tight if and only if $X$ is tight.*

**Proof.** For Part (i), fix $f \in C_b(\mathbb{D})$. Note that weak convergence implies both $\mathrm{E}^* f(X_n) \to \mathrm{E} f(X)$ and $\mathrm{E}_* f(X_n) = -\mathrm{E}^*[-f(X_n)] \to -\mathrm{E}[-f(X)] =$

$\mathrm{E}f(X)$, and the desired result follows since $f$ is arbitrary. For Part (ii), fix $\epsilon > 0$. Assume $X$ is tight, and choose a compact $K$ so that $\mathrm{P}(X \in K) \geq 1 - \epsilon$. By Part (ii) of the portmanteau theorem, $\liminf \mathrm{P}_*(X_n \in K^\delta) \geq \mathrm{P}(X \in K^\delta) \geq 1 - \epsilon$ for every $\delta > 0$. Hence $X_n$ is asymptotically tight. Now assume that $X_n$ is asymptotically tight, fix $\epsilon > 0$, and choose a compact $K$ so that $\liminf \mathrm{P}_*(X_n \in K^\delta) \geq 1 - \epsilon$ for every $\delta > 0$. By Part (iii) of the portmanteau theorem, $\mathrm{P}(X \in \overline{K^\delta}) \geq \limsup \mathrm{P}^*(X_n \in \overline{K^\delta}) \geq \liminf \mathrm{P}_*(X_n \in \overline{K^\delta}) \geq 1 - \epsilon$. By letting $\delta \downarrow 0$, we obtain that $X$ is tight.□

Prohorov's theorem (given below) tells us that asymptotic measurability and asymptotic tightness together almost gives us weak convergence. This "almost-weak-convergence" is *relative compactness*. A sequence $X_n$ is relatively compact if every subsequence $X_{n'}$ has a further subsequence $X_{n''}$ which converges weakly to a tight Borel law. Weak convergence happens when all of the limiting Borel laws are the same. Note that when all of the limiting laws assign probability one to a fixed Polish space, there is a converse of Prohorov's theorem, that relative compactness of $X_n$ implies asymptotic tightness of $X_n$ (and hence also uniform tightness). Details of this result are discussed in Chapter 1.12 of VW, but we do not pursue it further here.

THEOREM 7.13 *(Prohorov's theorem) If the sequence $X_n$ is asymptotically measurable and asymptotically tight, then it has a subsequence $X_{n'}$ that converges weakly to a tight Borel law.*

The proof, which we omit, is given in Chapter 1.3 of VW. Note that the conclusion of Prohorov's theorem does not state that $X_n$ is relatively compact, and thus it appears as if we have broken our earlier promise. However, if $X_n$ is asymptotically measurable and asymptotically tight, then every subsequence $X_{n'}$ is also asymptotically measurable and asymptotically tight. Thus repeated application of Prohorov's theorem does indeed imply relative compactness of $X_n$.

A natural question to ask at this juncture is: under what circumstances does asymptotic measurability and/or tightness of the marginal sequences $X_n$ and $Y_n$ imply asymptotic measurability and/or tightness of the joint sequence $(X_n, Y_n)$? This question is answered in the following lemma:

LEMMA 7.14 *Let $X_n : \Omega_n \mapsto \mathbb{D}$ and $Y_n : \Omega_n \mapsto \mathbb{E}$ be sequences of maps. Then the following are true:*

*(i) $X_n$ and $Y_n$ are both asymptotically tight if and only if the same is true for the joint sequence $(X_n, Y_n) : \Omega_n \mapsto \mathbb{D} \times \mathbb{E}$.*

*(ii) Asymptotically tight sequences $X_n$ and $Y_n$ are both asymptotically measurable if and only if $(X_n, Y_n) : \Omega_n \mapsto \mathbb{D} \times \mathbb{E}$ is asymptotically measurable.*

**Proof.** Let $d$ and $e$ be the metrics for $\mathbb{D}$ and $\mathbb{E}$, respectively, and let $\mathbb{D} \times \mathbb{E}$ be endowed with the product topology. Now note that a set in $\mathbb{D} \times \mathbb{E}$ of the form $K_1 \times K_2$ is compact if and only if $K_1$ and $K_2$ are both compact. Let $\pi_j K$, $j = 1, 2$, be the projections of $K$ onto $\mathbb{D}$ and $\mathbb{E}$, respectively. To be precise, the projection $\pi_1 K$ consists of all $x \in \mathbb{D}$ such that $(x, y) \in K$ for some $y \in \mathbb{E}$, and $\pi_2 K$ is analogously defined for $\mathbb{E}$. We leave it as an exercise to show that $\pi_1 K$ and $\pi_2 K$ are both compact. It is easy to see that $K$ is contained in $\pi_1 K \times \pi_2 K$. Using the product space metric $\rho((x_1, y_1), (x_2, y_2)) = d(x_1, x_2) \vee e(y_1, y_2)$ (one of several metrics generating the product topology), we now have for $K_1 \in \mathbb{D}$ and $K_2 \in \mathbb{E}$ that $(K_1 \times K_2)^\delta = K_1^\delta \times K_2^\delta$. Part (i) now follows from the definition of asymptotic tightness.

Let $\pi_1 : \mathbb{D} \times \mathbb{E} \mapsto \mathbb{D}$ be the projection onto the first coordinate, and note that $\pi_1$ is continuous. Thus for any $f \in C_b(\mathbb{D})$, $f \circ \pi_1 \in C_b(\mathbb{D} \times \mathbb{E})$. Hence joint asymptotic measurability of $(X_n, Y_n)$ implies asymptotic measurability of $X_n$ by the definition of asymptotic measurability. The same argument holds for $Y_n$. The difficult part of proving Part (ii) is the implication that asymptotic tightness plus asymptotic measurability of both marginal sequences yields asymptotic measurability of the joint sequence. We omit this part of the proof, but it can be found in Chapter 1.4 of VW.□

A very useful consequence of Lemma 7.14 is Slutsky's theorem. Note that the proof of Slutsky's theorem (given below) also utilizes both Prohorov's theorem and the continuous mapping theorem.

THEOREM 7.15 *(Slutsky's theorem) Suppose $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$, where $X$ is separable and $c$ is a fixed constant. Then the following are true:*

*(i)* $(X_n, Y_n) \rightsquigarrow (X, c)$.

*(ii) If $X_n$ and $Y_n$ are in the same metric space, then $X_n + Y_n \rightsquigarrow X + c$.*

*(iii) Assume in addition that the $Y_n$ are scalars. Then whenever $c \in \mathbb{R}$, $Y_n X_n \rightsquigarrow cX$. Also, whenever $c \neq 0$, $X_n / Y_n \rightsquigarrow X/c$.*

**Proof.** By completing the metric space for $X$, we can without loss of generality assume that $X$ is tight. Thus by Lemma 7.14, $(X_n, Y_n)$ is asymptotically tight and asymptotically measurable. Thus by Prohorov's theorem, all subsequences of $(X_n, Y_n)$ have further subsequences which converge to tight limits. Since these limit points have marginals $X$ and $c$, and since the marginals in this case completely determine the joint distribution, we have that all limiting distributions are uniquely determined as $(X, c)$. This proves Part (i). Parts (ii) and (iii) now follow from the continuous mapping theorem.□

### *7.2.2 Spaces of Bounded Functions*

Now we consider the setting where the $X_n$ are stochastic processes with index set $T$. The natural metric space for weak convergence in this setting is $\ell^\infty(T)$. A nice feature of this setting is the fact that asymptotic measurability of $X_n$ follows from asymptotic measurability of $X_n(t)$ for each $t \in T$:

LEMMA 7.16 *Let the sequence of maps $X_n$ in $\ell^\infty(T)$ be asymptotically tight. Then $X_n$ is asymptotically measurable if and only if $X_n(t)$ is asymptotically measurable for each $t \in T$.*

**Proof.** Let $f_t : \ell^\infty(T) \mapsto \mathbb{R}$ be the marginal projection at $t \in T$, i.e., $f_t(x) = x(t)$ for any $x \in \ell^\infty(T)$. Since each $f_t$ is continuous, asymptotic measurability of $X_n$ implies asymptotic measurability of $X_n(t)$ for each $t \in T$. Now assume that $X_n(t)$ is asymptotically measurable for each $t \in T$. Then Lemma 7.14 implies asymptotic measurability for all finite-dimensional marginals $(X_n(t_1), \ldots, X_n(t_k))$. Consequently, $f(X_n)$ is asymptotically measurable for all $f \in \mathcal{F}$, for the subset of $C_b(\mathbb{D})$ defined in the proof of Lemma 7.3 given above in Section 7.2, where $\mathbb{D} = \ell^\infty(T)$. Since $\mathcal{F}$ is an algebra that separates points in $\ell^\infty(T)$, asymptotic measurability of $X_n$ follows from Lemma 7.9.□

We now verify that convergence of finite dimensional distributions plus asymptotic tightness is equivalent to weak convergence in $\ell^\infty(T)$:

THEOREM 7.17 *The sequence $X_n$ converges to a tight limit in $\ell^\infty(T)$ if and only if $X_n$ is asymptotically tight and all finite-dimensional marginals converge weakly to limits. Moreover, if $X_n$ is asymptotically tight and all of its finite-dimensional marginals $(X_n(t_1), \ldots, X_n(t_k))$ converge weakly to the marginals $(X(t_1), \ldots, X(t_k))$ of a stochastic process $X$, then there is a version of $X$ such that $X_n \rightsquigarrow X$ and $X$ resides in $UC(T, \rho)$ for some semimetric $\rho$ making $T$ totally bounded.*

**Proof.** The result that "asymptotic tightness plus convergence of finite-dimensional distributions implies weak convergence" follows from Prohorov's theorem and Lemmas 7.9 and 7.1, using the vector lattice and subalgebra $\mathcal{F} \subset C_b(\mathbb{D})$ defined above in the proof of Lemma 7.3. The implication in the opposite direction follows easily from Lemma 7.12 and the continuous mapping theorem. Now assume $X_n$ is asymptotically tight and that all finite-dimensional distributions of $X_n$ converge to those of a stochastic process $X$. By asymptotic tightness of $X_n$, the probability that a version of $X$ lies in some $\sigma$-compact $K \subset \ell^\infty(T)$ is one. By Theorem 6.2, $K \subset UC(T, \rho)$ for some semimetric $\rho$ making $T$ totally bounded.□

Recall Theorem 2.1 and the Condition (2.6) from Chapter 2. When Condition (2.6) holds for every $\epsilon > 0$, then we say that the sequence $X_n$ is *asymptotically uniformly $\rho$-equicontinuous in probability.* We are now in a position to prove Theorem 2.1 on Page 15. Note that the statement of this

theorem is slightly informal: Conditions (i) and (ii) of the theorem actually imply that $X_n \rightsquigarrow X'$ in $\ell^\infty(T)$ for some tight version $X'$ of $X$. Recall that $\|x\|_T \equiv \sup_{t\in T} |x(t)|$.

**Proof of Theorem 2.1 (see Page 15).** First assume $X_n \rightsquigarrow X$ in $\ell^\infty(T)$, where $X$ is tight. Convergence of all finite-dimensional distributions follows from the continuous mapping theorem. Now by Theorem 6.2, $\mathrm{P}(X \in UC(T,\rho))$ for some semimetric $\rho$ making $T$ totally bounded. Hence for every $\eta > 0$, there exists some compact $K \subset UC(T,\rho)$ so that

$$\liminf_{n\to\infty} \mathrm{P}_*(X_n \in K^\delta) \geq 1-\eta, \quad \text{for all } \delta > 0. \tag{7.3}$$

Fix $\epsilon, \eta > 0$, and let the compact set $K$ satisfy (7.3). By Theorem 6.2, there exists a $\delta_0 > 0$ so that $\sup_{x\in K} \sup_{s,t:\rho(s,t)<\delta_0} |x(s) - x(t)| \leq \epsilon/3$. Now

$$\begin{aligned}
&\mathrm{P}^* \left[ \sup_{s,t\in T:\rho(s,t)<\delta_0} |X_n(s) - X_n(t)| > \epsilon \right] \\
&\quad \leq \quad \mathrm{P}^* \left[ \sup_{s,t\in T:\rho(s,t)<\delta_0} |X_n(s) - X_t(t)| > \epsilon,\ X_n \in K^{\epsilon/3} \right] + \mathrm{P}^*(X_n \notin K^{\epsilon/3}) \\
&\quad \equiv \quad E_n
\end{aligned}$$

satisfies $\limsup_{n\to\infty} E_n \leq \eta$, since if $x \in K^{\epsilon/3}$ then $\sup_{s,t\in T:\rho(s,t)<\delta_0} |x(s) - x(t)| < \epsilon$. Thus $X_n$ is asymptotically uniformly $\rho$-continuous in probability since $\epsilon$ and $\eta$ were arbitrary.

Now assume that Conditions (i) and (ii) of the theorem hold. Lemma 7.18 below, the proof of which is given in Section 7.4, yields that $X_n$ is asymptotically tight. Thus the desired weak converge of $X_n$ follows from Theorem 7.17 above.□

LEMMA 7.18 *Assume Conditions (i) and (ii) of Theorem 2.1 hold. Then $X_n$ is asymptotically tight.*

The proof of Theorem 2.1 verifies that whenever $X_n \rightsquigarrow X$ and $X$ is tight, any semimetric $\rho$ defining a $\sigma$-compact set $UC(T,\rho)$ such that $\mathrm{P}(X \in UC(T,\rho)) = 1$ will also result in $X_n$ being uniformly $\rho$-equicontinuous in probability. What is not clear at this point is the converse, that any semimetric $\rho_*$ which enables uniform asymptotic equicontinuity of $X_n$ will also define a $\sigma$-compact set $UC(T,\rho_*)$ wherein $X$ resides with probability 1. The following theorem shows that, in fact, any semimetric which works for one of these implications will work for the other:

THEOREM 7.19 *Assume $X_n \rightsquigarrow X$ in $\ell^\infty(T)$, and let $\rho$ be a semimetric making $(T,\rho)$ totally bounded. Then the following are equivalent:*

*(i) $X_n$ is asymptotically uniformly $\rho$-equicontinuous in probability.*

*(ii)* $\mathrm{P}(X \in UC(T,\rho)) = 1$.

**Proof.** If we assume (ii), then (i) will follow by arguments given in the proof of Theorem 2.1 above. Now assume (i). For $x \in \ell^\infty(T)$, define $M_\delta(x) \equiv \sup_{s,t \in T:\rho(s,t)<\delta} |x(s) - x(t)|$. Note that if we restrict $\delta$ to $(0,1)$ then $x \mapsto M_{(\cdot)}(x)$, as a map from $\ell^\infty(T)$ to $\ell^\infty((0,1))$, is continuous since $|M_\delta(x) - M_\delta(y)| \leq 2\|x-y\|_T$ for all $\delta \in (0,1)$. Hence $M_{(\cdot)}(X_n) \rightsquigarrow M_{(\cdot)}(X)$ in $\ell^\infty((0,1))$. Condition (i) now implies that there exists a positive sequence $\delta_n \downarrow 0$ so that $\mathrm{P}^*(M_{\delta_n}(X_n) > \epsilon) \to 0$ for every $\epsilon > 0$. Hence $M_{\delta_n}(X) \rightsquigarrow 0$. This implies (ii) since $X$ is tight by Theorem 2.1.□

An interesting consequence of Theorems 2.1 and 7.19, in conjunction with Lemma 7.4, happens when $X_n \rightsquigarrow X$ in $\ell^\infty(T)$ and $X$ is a tight Gaussian process. Recall from Section 7.1 the semimetric $\rho_p(s,t) \equiv (E|X(s) - X(t)|^p)^{1/(p \vee 1)}$, for any $p \in (0,\infty)$. Then for any $p \in (0,\infty)$, $(T, \rho_p)$ is totally bounded, the sample paths of $X$ are $\rho_p$-continuous, and, furthermore, $X_n$ is asymptotically uniformly $\rho_p$-equicontinuous in probability. While any value of $p \in (0,\infty)$ will work, the choice $p = 2$ (the "standard deviation" metric) is often the most convenient to work with.

We now point out an equivalent condition for $X_n$ to be asymptotically uniformly $\rho$-equicontinuous in probability. This new condition, which is expressed in the following lemma, is sometimes easier to verify for certain settings (one of which occurs in the next chapter):

LEMMA 7.20 *Let $X_n$ be a sequence of stochastic processes indexed by $T$. Then the following are equivalent:*

*(i) There exists a semimetric $\rho$ making $T$ totally bounded and for which $X_n$ is uniformly $\rho$-equicontinuous in probability.*

*(ii) For every $\epsilon, \eta > 0$, there exists a finite partition $T = \cup_{i=1}^k T_i$ such that*

$$\limsup_{n\to\infty} \mathrm{P}^* \left( \sup_{1 \leq i \leq k} \sup_{s,t \in T_i} |X_n(s) - X_n(t)| > \epsilon \right) \quad < \quad \eta. \tag{7.4}$$

The proof is given in Section 7.4 below.

## 7.3 Other Modes of Convergence

Recall the definitions of convergence in probability and outer almost surely, for arbitrary maps $X_n : \Omega \mapsto \mathbb{D}$, as defined in Chapter 2. We now introduce two additional modes of convergence which can be useful in some settings. $X_n$ *converges almost uniformly* to $X$ if, for every $\epsilon > 0$, there exists a measurable set $A$ such that $\mathrm{P}(A) \geq 1 - \epsilon$ and $d(X_n, X) \to 0$ uniformly on $A$. $X_n$ *converges almost surely* to $X$ if $\mathrm{P}_*(\lim_{n\to\infty} d(X_n, X) = 0) = 1$. Note that an important distinction between almost sure and outer almost sure convergence is that, in the latter mode, there must exist a measurable majorant of $d(X_n, X)$ which goes to zero. This distinction is quite important

because it can be shown that almost sure convergence does not in general imply convergence in probability when $d(X_n, X)$ is not measurable. For this reason, we do not use the almost sure convergence mode in this book except rarely. One of those rare times is in Exercise 7.5.7, another is in Proposition 7.22 below which will be used in Chapter 10. The following lemma characterizes the relationships among the three remaining modes:

LEMMA 7.21 *Let $X_n, X : \Omega \mapsto \mathbb{D}$ be maps with $X$ 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* $X_{n''}$ *such that* $X_{n''} \xrightarrow{\text{as*}} X$.

*(iii)* $X_n \xrightarrow{\text{as*}} X$ *if and only if* $X_n$ *converges almost uniformly to* $X$ *if and only if* $\sup_{m \geq n} d(X_m, X) \xrightarrow{\text{P}} 0$.

The proof is given Section 7.4. Since almost uniform convergence and outer almost sure convergence are equivalent for sequences, we will not use the almost uniform mode very much.

The following proposition gives a connection between almost sure convergence and convergence in probability. We need this proposition for a continuous mapping result for bootstrapped processes presented in Chapter 10:

PROPOSITION 7.22 *Let $X_n, Y_n : \Omega \mapsto \mathbb{D}$ be maps with $Y_n$ measurable. Suppose every subsequence $n'$ has a further subsequence $n''$ such that $X_{n''} \to 0$ almost surely. Suppose also that $d(X_n, Y_n) \xrightarrow{\text{P}} 0$. Then $X_n \xrightarrow{\text{P}} 0$.*

**Proof.** For every subsequence $n'$ there exists a further subsequence $n''$ such that both $X_{n''} \to 0$ and $d(X_{n''}, Y_{n''})^* \to 0$ almost surely for some versions $d(X_{n''}, Y_{n''})^*$. Since $d(Y_n, 0) \leq d(X_n, 0) + d(X_n, Y_n)^*$, we have that $Y_{n''} \to 0$ almost surely. But this implies $Y_{n''} \xrightarrow{\text{as*}} 0$ since the $Y_n$ are measurable. Since the subsequence $n'$ was arbitrary, we now have that $Y_n \xrightarrow{\text{P}} 0$. Thus $X_n \xrightarrow{\text{P}} 0$ since $d(X_n, 0) \leq d(Y_n, 0) + d(X_n, Y_n)$.□

The next lemma describes several important relationships between weak convergence and convergence in probability. Before presenting it, we need to extend the definition of convergence in probability, in the setting where the limit is a constant, to allow the probability spaces involved to change with $n$ as is already permitted for weak convergence. We denote this modified convergence $X_n \xrightarrow{\text{P}} c$, and distinguish it from the previous form of convergence in probability only by context.

LEMMA 7.23 *Let $X_n, Y_n : \Omega_n \mapsto \mathbb{D}$ be maps, $X : \Omega \mapsto \mathbb{D}$ be Borel measurable, and $c \in \mathbb{D}$ be a constant. Then*

*(i) If* $X_n \rightsquigarrow X$ *and* $d(X_n, Y_n) \xrightarrow{\text{P}} 0$, *then* $Y_n \rightsquigarrow X$.

*(ii)* $X_n \xrightarrow{\mathrm{P}} X$ *implies* $X_n \rightsquigarrow X$.

*(iii)* $X_n \xrightarrow{\mathrm{P}} c$ *if and only if* $X_n \rightsquigarrow c$.

**Proof.** We first prove (i). Let $F \subset \mathbb{D}$ be closed, and fix $\epsilon > 0$. Then $\limsup_{n\to\infty} \mathrm{P}^*(Y_n \in F) = \limsup_{n\to\infty} \mathrm{P}^*(Y_n \in F, d(X_n, Y_n)^* \leq \epsilon) \leq \limsup_{n\to\infty} \mathrm{P}^*(X_n \in \overline{F^\epsilon}) \leq P(X \in \overline{F^\epsilon})$. The result follows by letting $\epsilon \downarrow 0$. Now assume $X_n \xrightarrow{\mathrm{P}} X$. Since $X \rightsquigarrow X$, $d(X, X_n) \xrightarrow{\mathrm{P}} 0$ implies $X_n \rightsquigarrow X$ by (i), thus (ii) follows. We now prove (iii). $X_n \xrightarrow{\mathrm{P}} c$ implies $X_n \rightsquigarrow c$ by (ii). Now assume $X_n \rightsquigarrow c$, and fix $\epsilon > 0$. Note that $\mathrm{P}^*(d(X_n, c) \geq \epsilon) = \mathrm{P}^*(X_n \notin B(c, \epsilon))$, where $B(c, \epsilon)$ is the open $\epsilon$-ball around $c \in \mathbb{D}$. By the portmanteau theorem, $\limsup_{n\to\infty} \mathrm{P}^*(X_n \notin B(c, \epsilon)) \leq \mathrm{P}(X \notin B(c, \epsilon)) = 0$. Thus $X_n \xrightarrow{\mathrm{P}} c$ since $\epsilon$ is arbitrary, and (iii) follows.□

We now present a generalized continuous mapping theorem that allows for sequences of maps $g_n$ which converge to $g$ in a fairly general sense. In the exercises at the end of the chapter, we consider an instance of this where one is interested in maximizing a stochastic process $\{X_n(t), t \in T\}$ over an "approximation" $T_n$ of a subset $T_0 \subset T$. As a specific motivation, suppose $T$ is high dimensional. The computational burden of computing the supremum of $X_n(t)$ over $T$ may be reduced by choosing a finite mesh $T_n$ which closely approximates $T$. In what follows, $(\mathbb{D}, d)$ and $(\mathbb{E}, e)$ are metric spaces.

THEOREM 7.24 *(Extended continuous mapping). Let* $\mathbb{D}_n \subset \mathbb{D}$ *and* $g_n : \mathbb{D}_n \mapsto \mathbb{E}$ *satisfy the following: if* $x_n \to x$ *with* $x_n \in \mathbb{D}_n$ *for all* $n \geq 1$ *and* $x \in \mathbb{D}_0$, *then* $g_n(x_n) \to g(x)$, *where* $\mathbb{D}_0 \subset \mathbb{D}$ *and* $g : \mathbb{D}_0 \mapsto \mathbb{E}$. *Let* $X_n$ *be maps taking values in* $\mathbb{D}_n$, *and let* $X$ *be Borel measurable and separable with* $\mathrm{P}_*(X \in \mathbb{D}_0) = 1$. *Then*

*(i)* $X_n \rightsquigarrow X$ *implies* $g_n(X_n) \rightsquigarrow g(X)$.

*(ii)* $X_n \xrightarrow{\mathrm{P}} X$ *implies* $g_n(X_n) \xrightarrow{\mathrm{P}} g(X)$.

*(iii)* $X_n \xrightarrow{\mathrm{as}^*} X$ *implies* $g_n(X_n) \xrightarrow{\mathrm{as}^*} g(X)$.

The proof can be found in Chapter 1.11 of VW, and we omit it here. It is easy to see that if $g : \mathbb{D} \mapsto \mathbb{E}$ is continuous at all points in $\mathbb{D}_0$, and if we set $g_n = g$ and $\mathbb{D}_n = \mathbb{D}$ for all $n \geq 1$, then the standard continuous mapping theorem (Theorem 7.7), specialized to the setting where $X$ is separable, is a corollary of Part (i) of the above theorem.

The following theorem gives another kind of continuous mapping result for sequences which converge in probability and outer almost surely. When $X$ is separable, the conclusions of this theorem are a simple corollary of Theorem 7.24.

THEOREM 7.25 *Let* $g : \mathbb{D} \mapsto \mathbb{E}$ *be continuous at all points in* $\mathbb{D}_0 \subset \mathbb{D}$, *and let* $X$ *be Borel measurable with* $\mathrm{P}_*(X \in \mathbb{D}_0) = 1$. *Then*

*(i)* $X_n \xrightarrow{\mathrm{P}} X$ *implies* $g(X_n) \xrightarrow{\mathrm{P}} g(X)$.

*(ii)* $X_n \xrightarrow{\mathrm{as*}} X$ *implies* $g(X_n) \xrightarrow{\mathrm{as*}} g(X)$.

**Proof.** Assume $X_n \xrightarrow{\mathrm{P}} X$, and fix $\epsilon > 0$. Define $B_k$ to be all $x \in \mathbb{D}$ such that the $1/k$-ball around $x$ contains points $y$ and $z$ with $e(g(y), g(z)) > \epsilon$. Part of the proof of Theorem 7.7 in Section 7.4 verifies that $B_k$ is open. It is clear that $B_k$ decreases as $k$ increases. Furthermore, $\mathrm{P}(X \in B_k) \downarrow 0$, since every point in $\cap_{k=1}^{\infty} B_k$ is a point of discontinuity of $g$. Now the outer probability that $e(g(X_n), g(X)) > \epsilon$ is bounded above by the outer probability that either $X \in B_k$ or $d(X_n, X) \geq 1/k$. But this last outer probability converges to $\mathrm{P}^*(X \in B_k)$ since $d(X_n, X) \xrightarrow{\mathrm{P}} 0$. Part (i) now follows by letting $k \to \infty$ and noting that $\epsilon$ was arbitrary. Assume that $X_n \xrightarrow{\mathrm{as*}} X$. Note that a minor modification of the proof of Part (i) verifies that $\sup_{m \geq n} d(X_m, X) \xrightarrow{\mathrm{P}} 0$ implies $\sup_{m \geq n} e(g(X_n), g(X)) \xrightarrow{\mathrm{P}} 0$. Now Part (iii) of Lemma 7.21 yields that $X_n \xrightarrow{\mathrm{as*}} X$ implies $g(X_n) \xrightarrow{\mathrm{as*}} g(X)$.□

We now present a useful outer almost sure representation result for weak convergence. Such representations allow the conversion of certain weak convergence problems into problems about convergence of fixed sequences. We give an illustration of this approach in the proof of Proposition 7.27 below.

THEOREM 7.26 *Let* $X_n : \Omega_n \mapsto \mathbb{D}$ *be a sequence of maps, and let* $X_\infty$ *be Borel measurable and separable. If* $X_n \rightsquigarrow X_\infty$, *then there exists a probability space* $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{P})$ *and maps* $\tilde{X}_n : \tilde{\Omega} \mapsto \mathbb{D}$ *with*

*(i)* $\tilde{X}_n \xrightarrow{\mathrm{as*}} \tilde{X}_\infty$;

*(i)* $\mathrm{E}^* f(\tilde{X}_n) = \mathrm{E}^* f(X_n)$, *for every bounded* $f : \mathbb{D} \mapsto \mathbb{R}$ *and all* $1 \leq n \leq \infty$.

*Moreover,* $\tilde{X}_n$ *can be chosen to be equal to* $X_n \circ \phi_n$, *for all* $1 \leq n \leq \infty$, *where the* $\phi_n : \tilde{\Omega} \mapsto \Omega_n$ *are measurable and perfect maps and* $P_n = \tilde{P} \circ \phi_n^{-1}$.

The proof can be found in Chapter 1.10 of VW, and we omit it here. Recall the definition of perfect maps from Chapter 6. In the setting of the above theorem, if the $\tilde{X}_n$ are constructed from the perfect maps $\phi_n$, then $\left[f(\tilde{X}_n)\right]^* = [f(X_n)]^* \circ \phi_n$ for all bounded $f : \mathbb{D} \mapsto \mathbb{R}$. Thus the equivalence between $\tilde{X}_n$ and $X_n$ can be made much stronger than simply equivalence in law.

The following proposition can be useful in studying weak convergence of certain statistics which can be expressed as stochastic integrals. For example, the Wilcoxon statistic can be expressed in this way. The proof of the proposition provides the illustration of Theorem 7.26 promised above.

PROPOSITION 7.27 *Let* $X_n, G_n \in D[a, b]$ *be stochastic processes with* $X_n \rightsquigarrow X$ *and* $G_n \xrightarrow{\mathrm{P}} G$ *in* $D[a, b]$, *where* $X$ *is bounded with continuous*

*sample paths, $G$ is fixed, and $G_n$ and $G$ have total variation bounded by $K < \infty$. Then $\int_a^{(\cdot)} X_n(s)dG_n(s) \rightsquigarrow \int_a^{(\cdot)} X(s)dG(s)$ in $D[a,b]$.*

**Proof.** First, Slutsky's theorem and Lemma 7.23 establish that $(X_n, G_n) \rightsquigarrow (X, G)$. Next, Theorem 7.26 tells us that there exists a new probability space and processes $\tilde{X}_n$, $\tilde{X}$, $\tilde{G}_n$ and $\tilde{G}$ which have the same outer integrals for bounded functions as $X_n$, $X$, $G_n$ and $G$, respectively, but which also satisfy $(\tilde{X}_n, \tilde{G}_n) \stackrel{\text{as*}}{\to} (\tilde{X}, \tilde{G})$. By the continuity of the sample paths of $\tilde{X}$, we have for each integer $m \geq 1$, that there exists a finite partition of $[a,b]$, $a = t_0 < t_1 < \cdots < t_{k_m} = b$, with $\mathrm{P}(M_m > 1/m) \leq 1/m$, where

$$M_m \equiv \max_{1 \leq j \leq k_m} \sup_{s,t \in (t_{j-1}, t_j]} |\tilde{X}(s) - \tilde{X}(t)|.$$

Define $\tilde{X}_m \in D[a,b]$ such that $\tilde{X}_m(a) = \tilde{X}(a)$ and $\tilde{X}_m(t) \equiv \sum_{j=1}^m 1\{t_{j-1} < t \leq t_j\}\tilde{X}(t_j)$, for $t \in (a,b]$. Note that for integrals over the range $(a,t]$, for $t \in [a,b]$, we define the value of the integral to be zero when $t = a$ since $(a,a]$ is the null set. We now have, for any $t \in [a,b]$, that

$$\begin{aligned}
&\left| \int_a^t \tilde{X}_n(s)d\tilde{G}_n(s) - \int_a^t \tilde{X}(s)d\tilde{G}(s) \right| \\
&\leq \int_a^b \left| \tilde{X}_n(s) - \tilde{X}(s) \right| \times |d\tilde{G}_n(s)| + \int_a^b \left| \tilde{X}_m(s) - \tilde{X}(s) \right| \times |d\tilde{G}_n(s)| \\
&\quad + \left| \int_a^t \tilde{X}_m(s) \left\{ d\tilde{G}_n(s) - d\tilde{G}(s) \right\} \right| + \int_0^t \left| \tilde{X}_m(s) - \tilde{X}(s) \right| \times |d\tilde{G}(s)| \\
&\leq K \left( \|\tilde{X}_n - \tilde{X}\|_{[a,b]} + 2M_m \right) \\
&\quad + \left| \sum_{j=1}^m \tilde{X}(t_j) \int_{(t_{j-1}, t_j] \cap (a,t]} \left\{ d\tilde{G}_n(s) - d\tilde{G}(s) \right\} \right| \\
&\leq K \left( \|\tilde{X}_n - \tilde{X}\|^*_{[a,b]} + 2M_m \right) + k_m \left( \|\tilde{X}\|_{[a,b]} \times \|\tilde{G}_n - \tilde{G}\|^*_{[a,b]} \right) \\
&\equiv E_n(m).
\end{aligned}$$

Note that $E_n(m)$ is measurable and $\to 0$ almost surely. Define $D_n$ to be the infimum of $E_n(m)$ over all integers $m \geq 1$. Since $D_n \stackrel{\text{as*}}{\to} 0$ and $D_n$ is measurable, we have that $\int_a^{(\cdot)} \tilde{X}_n(s)d\tilde{G}_n(s) \stackrel{\text{as*}}{\to} \int_a^{(\cdot)} \tilde{X}(s)d\tilde{G}(s)$. Choose any $f \in C_b(D[a,b])$, and note that the map $(x,y) \mapsto f\left( \int_a^{(\cdot)} x(s)dy(s) \right)$, for $x, y \in D[a,b]$ with the total variation of $y$ bounded, is bounded. Thus

$$\begin{aligned} \mathrm{E}^* f\left(\int_a^{(\cdot)} X_n(s)dG_n(s)\right) &= \mathrm{E}^* f\left(\int_a^{(\cdot)} \tilde{X}_n(s)d\tilde{G}_n(s)\right) \\ &\to \mathrm{E} f\left(\int_a^{(\cdot)} \tilde{X}(s)d\tilde{G}(s)\right) \\ &= \mathrm{E} f\left(\int_a^{(\cdot)} X(s)dG(s)\right). \end{aligned}$$

Since this convergence holds for all $f \in C_b(D[a,b])$, the desired result now follows.□

We give one more result before closing this chapter. The result applies to certain weak convergence settings involving questions that are easier to answer for measurable maps. The following lemma shows that a nonmeasurable, weakly convergent sequence $X_n$ is usually quite close to a measurable sequence $Y_n$:

LEMMA 7.28 *Let $X_n : \Omega_n \mapsto \mathbb{D}$ be a sequence of maps. If $X_n \rightsquigarrow X$, where $X$ is Borel measurable and separable, then there exists a Borel measurable sequence $Y_n : \Omega_n \mapsto \mathbb{D}$ with $d(X_n, Y_n) \xrightarrow{\mathrm{P}} 0$.*

The proof can be found in Chapter 1.10 of VW, and we omit it here.

## 7.4 Proofs

**Proof of Lemma 7.1.** Clearly (i) implies (ii). Now assume (ii). For every open $G \subset \mathbb{D}$, define the sequence of functions $f_m(x) = [md(x, \mathbb{D}-G)] \wedge 1$, for integers $m \geq 1$, and note that each $f_m$ is bounded and Lipschitz continuous and $f_m \uparrow 1\{G\}$ as $m \to \infty$. By monotone convergence, $L_1(G) = L_2(G)$. Since this is true for every open $G \subset \mathbb{D}$, including $G = \mathbb{D}$, the collection of Borel sets for which $L_1(B) = L_2(B)$ is a $\sigma$-field and is at least as large as the Borel $\sigma$-field. Hence (ii) implies (i). The equivalence of (i) and (iii) under separability follows from Theorem 1.12.2 of VW and we omit the details here.

The fact that (i) implies (iv) is obvious. Now assume $L_1$ and $L_2$ are tight and that (iv) holds. Fix $\epsilon > 0$, and choose a compact $K \subset \mathbb{D}$ such that $L_1(K) \wedge L_2(K) \geq 1 - \epsilon$. According to a version of the Stone-Weierstrass theorem given in Jameson (1974, p. 263), a vector lattice $\mathcal{F} \subset C_b(K)$ that includes the constants and separates points of $K$ is uniformly dense in $C_b(K)$. Choose a $g \in C_b(\mathbb{D})$ for which $0 \leq g \leq 1$, and select an $f \in \mathcal{F}$ such that $\sup_{x \in K} |g(x) - f(x)| \leq \epsilon$. Now we have $\left|\int g dL_1 - \int g dL_2\right| \leq \left|\int_K g dL_1 - \int_K g dL_2\right| + 2\epsilon \leq \left|\int_K (f \wedge 1)^+ dL_1 - \int_K (f \wedge 1)^+ dL_2\right| + 4\epsilon = 4\epsilon$. The last equality follows since $(f \wedge 1)^+ \in \mathcal{F}$. Thus $\int g dL_1 = \int g dL_2$ since $\epsilon$ is arbitrary. By adding and subtracting scalars, we can verify that the same result holds for all $g \in C_b(\mathbb{D})$. Hence (i) holds.□

**Proof of Lemma 7.2.** The equivalence of (i) and (ii) is an immediate consequence of Theorem 6.2. Now assume (ii) holds. Then $|X|$ is bounded almost surely. Hence for any pair of sequences $s_n, t_n \in T$ such that $\rho_0(s_n, t_n) \to 0$, $X(s_n) - X(t_n) \xrightarrow{\mathrm{P}} 0$. Thus $X \in UC(T, \rho_0)$ with probability 1. It remains to show that $(T, \rho_0)$ is totally bounded. Let the pair of sequences $s_n, t_n \in T$ satisfy $\rho(s_n, t_n) \to 0$. Then $X(s_n) - X(t_n) \xrightarrow{\mathrm{P}} 0$ and thus $\rho_0(s_n, t_n) \to 0$. This means that since $(T, \rho)$ is totally bounded, we have for every $\epsilon > 0$ that there exists a finite $T_\epsilon \subset T$ such that $\sup_{t \in T} \inf_{s \in T_\epsilon} \rho_0(s, t) < \epsilon$. Thus $(T, \rho_0)$ is also totally bounded, and the desired result follows.□

**Proof of Theorem 7.6.** Assume (i), and note that (i) implies (vii) trivially. Now assume (vii), and fix an open $G \subset \mathbb{D}$. As in the proof of Lemma 7.1 above, there exists a sequence of nonnegative, Lipschitz continuous functions $f_m$ with $0 \le f_m \uparrow 1\{G\}$. Now for each integer $m \ge 1$, $\liminf \mathrm{P}_*(X_n \in G) \ge \liminf \mathrm{E}_* f_m(X_n) = \mathrm{E} f_m(X)$. Taking the limit as $m \to \infty$ yields (ii). Thus (vii)⇒(ii). The equivalence of (ii) and (iii) follows by taking complements.

Assume (ii) and let $f$ be lower semicontinuous with $f \ge 0$. Define the sequence of functions $f_m = \sum_{i=1}^{m^2} (1/m) 1\{G_i\}$, where $G_i = \{x : f(x) > i/m\}$, $i = 1, \ldots, m^2$. Thus $f_m$ "rounds" $f$ down to $i/m$ if $f(x) \in (i/m, (i+1)/m]$ for any $i = 0, \ldots, m^2 - 1$ and $f_m = m$ when $f(x) > m$. Hence $0 \le f_m \le f \wedge m$ and $|f_m - f|(x) \le 1/m$ whenever $f(x) \le m$. Fix $m$. Note that each $G_i$ is open by the definition of lower semicontinuity. Thus

$$\begin{aligned}
\liminf_{n \to \infty} \mathrm{E}_* f(X_n) &\ge \liminf_{n \to \infty} \mathrm{E}_* f_m(X_n) \\
&\ge \sum_{i=1}^{m^2} (1/m) \left[ \liminf_{n \to \infty} \mathrm{P}_*(X_n \in G_i) \right] \\
&\ge \sum_{i=1}^{m^2} (1/m) \mathrm{P}(X \in G_i) \\
&= \mathrm{E} f_m(X).
\end{aligned}$$

Thus (ii) implies (iv) after letting $m \to \infty$ and adding then subtracting a constant as needed to compensate for the lower bound of $f$. The equivalence of (iv) and (v) follows by replacing $f$ with $-f$. Assume (v) (and thus also (iv)). Since a continuous function is both upper and lower semicontinuous, we have for any $f \in C_b(\mathbb{D})$ that $\mathrm{E} f(X) \ge \limsup \mathrm{E}^* f(X_n) \ge \liminf \mathrm{E}_* f(X_n) \ge \mathrm{E} f(X)$. Hence (v) implies (i).

Assume (ii) (and hence also (iii)). For any Borel set $B \subset \mathbb{D}$, $L(B^\circ) \le \liminf_{n \to \infty} \mathrm{P}_*(X_n \in B^\circ) \le \limsup_{n \to \infty} \mathrm{P}^*(X_n \in \overline{B}) \le L(\overline{B})$; however, the forgoing inequalities all become equalities when $L(\delta B) = 0$. Thus (ii) implies (vi). Assume (vi), and let $F$ be closed. For each $\epsilon > 0$ define $F^\epsilon = \{x : d(x, F) < \epsilon\}$. Since the sets $\delta F^\epsilon$ are disjoint, $L(\delta F^\epsilon) > 0$

for at most countably many $\epsilon$. Hence we can choose a sequence $\epsilon_m \downarrow 0$ so that $L(\delta F^{\epsilon_n}) = 0$ for each integer $m \geq 1$. Note that for fixed $m$, $\limsup_{n\to\infty} \mathrm{P}^*(X_n \in F) \leq \limsup_{n\to\infty} \mathrm{P}^*(X_n \in \overline{F^{\epsilon_m}}) = L(\overline{F^{\epsilon_m}})$. By letting $m \to \infty$, we obtain that (vi) implies (ii). Thus Conditions (i)–(vii) are all equivalent.

The equivalence of (i)–(vii) to (viii) when $L$ is separable follows from Theorem 1.12.2 of VW, and we omit the details.□

**Proof of Theorem 7.7.** The set of all points at which $g$ is not continuous can be expressed as $D_g \equiv \bigcup_{m=1}^{\infty} \bigcap_{k=1}^{\infty} G_k^m$, where $G_k^m$ consists of all $x \in \mathbb{D}$ so that $e(g(y), g(z)) > 1/m$ for some $y, z \in B_{1/k}(x)$, where $e$ is the metric for $\mathbb{E}$. Note that the complement of $G_k^m$, $(G_k^m)^c$, consists of all $x$ for which $e(g(y), g(z)) \leq 1/m$ for all $y, z \in B_{1/k}(x)$. Now if $x_n \to x$, then for any $y, z \in B_{1/k}(x)$, we have that $y, z \in B_{1/k}(x_n)$ for all $n$ large enough. Hence $(G_k^m)^c$ is closed and thus $G_k^m$ is open. This means that $D_g$ is a Borel set. Let $F \subset \mathbb{E}$ be closed, and let $\{x_n\} \in g^{-1}(F)$ be a sequence for which $x_n \to x$. If $x$ is a continuity point of $g$, then $x \in g^{-1}(F)$. Otherwise $x \in D_g$. Hence $\overline{g^{-1}(F)} \subset g^{-1}(F) \cup D_g$. Since $g$ is continuous on the range of $X$, there is a version of $g(X)$ that is Borel Measurable. By the portmanteau theorem, $\limsup \mathrm{P}^*(g(X_n) \in F) \leq \limsup \mathrm{P}^*(X_n \in \overline{g^{-1}(F)}) \leq \mathrm{P}(X \in \overline{g^{-1}(F)})$. Since $D_g$ has probability zero under the law of $X$, $\mathrm{P}(X \in \overline{g^{-1}(F)}) = \mathrm{P}(g(X) \in F)$. Reapplying the portmanteau theorem, we obtain the desired result.□

**Proof of Lemma 7.10.** It is easy to see that uniform tightness implies asymptotic tightness. To verify equivalence going the other direction, assume that $X_n$ is tight for each $n \geq 1$ and that the sequence is asymptotically tight. Fix $\epsilon > 0$, and choose a compact $K_0$ for which $\liminf \mathrm{P}(X_n \in K_0^{\delta}) \geq 1-\epsilon$ for all $\delta > 0$. For each integer $m \geq 1$, choose an $n_m < \infty$ so that $\mathrm{P}(X_n \in K_0^{1/m}) \geq 1 - 2\epsilon$ for all $n \geq n_m$. For each integer $n \in (n_m, n_{m+1}]$, choose a compact $\tilde{K}_n$ so that $\mathrm{P}(X_n \in \tilde{K}_n) \geq 1 - \epsilon/2$ and an $\eta_n \in (0, 1/m)$ so that $\mathrm{P}\left(X_n \in K_0^{1/m} - \overline{K_0^{\eta_n}}\right) < \epsilon/2$. Let $K_n = (\tilde{K}_n \cup K_0) \cap \overline{K_0^{\eta_n}}$, and note that $K_0 \subset K_n \subset K_0^{1/m}$ and that $K_n$ is compact. We leave it as an exercise to show that $K \equiv \bigcup_{n=1}^{\infty} K_n$ is also compact. Now $\mathrm{P}(X_n \in K_n) \geq 1 - 3\epsilon$ for all $n \geq 1$, and thus $\mathrm{P}(X_n \in K) \geq 1 - 3\epsilon$ for all $n \geq 1$. Uniform tightness follows since $\epsilon$ was arbitrary.□

**Proof of Lemma 7.18.** Fix $\zeta > 0$. The conditions imply that $\mathrm{P}^*(\|X_n\|_T^* > M) < \zeta$ for some $M < \infty$. Let $\epsilon_m$ be a positive sequence converging down to zero and let $\eta_m \equiv 2^{-m}\zeta$. By Condition (ii), there exists a positive sequence $\delta_m \downarrow 0$ so that

$$\limsup_{n\to\infty} \mathrm{P}^* \left( \sup_{s,t \in T : \rho(s,t) < \delta_m} |X_n(s) - X_n(t)| > \epsilon_m \right) < \eta_m.$$

Now fix $m$. By the total boundedness of $T$, there exists a finite set of disjoint partions $T_1, \ldots, T_k$ so that $T = \cup_{i=1}^{k} T_i$ and so that

$$\mathrm{P}^*\left(\max_{1\le i\le k}\sup_{s,t\in T_i}|X_n(s)-X_n(t)|>\epsilon_m\right)<\eta_m.$$

Let $z_1,\ldots,z_p$ be the set of all functions on $\ell^\infty(T)$ which are constant on each $T_i$ and which take only the values $\pm i\epsilon_m$, for $i=0,\ldots,K$, where $K$ is the largest integer $\le M/\epsilon_m$. Now let $K_m$ be the union of the closed balls of radius $\epsilon_m$ around the $z_i$, $i=1,\ldots,p$. This construction ensures that if $\|x\|_T\le M$ and $\max_{1\le i\le k}\sup_{s,t\in T_i}|x(s)-x(t)|\le\epsilon_m$, then $x\in K_m$. This construction can be repeated for each $m\ge 1$.

Let $K=\cap_{m=1}^\infty K_m$, and note that $K$ is totally bounded and closed. Total boundedness follows since each $K_m$ is a union of finite $\epsilon_m$ balls which cover $K$ and $\epsilon_m\downarrow 0$. We leave the proof that $K$ is closed as an exercise. Now we show that for every $\delta>0$, there is an $m<\infty$ so that $K^\delta\supset\cap_{i=1}^m K_i$. If this were not true, then there would be a sequence $\{z_m\}\notin K^\delta$ with $z_m\in\cap_{i=1}^m K_i$ for every $m\ge 1$. This sequence has a subsequence $\{z_{m_1(k)},k\ge 1\}$ contained in one of the open balls making up $K_1$. This subsequence has a further subsequence $\{z_{m_2(k)},k\ge 1\}$ contained in one of the open balls making up $K_2$. We can continue with this process to generate, for each integer $j\ge 1$, a subsequence $\{z_{m_j(k)},k\ge 1\}$ contained in the intersection $\cap_{i=1}^j B_i$, where each $B_i$ is one of the open balls making up $K_i$. Define a new subsequence $\tilde z_k=z_{m_k(k)}$, and note that $\tilde z_k$ is Cauchy with limit in $K$ since $K$ is closed. However, this contradicts $d(z_m,K)\ge\delta$ for all $m$. Hence the complement of $K^\delta$ is contained in the complement of $\cap_{i=1}^m K_i$ for some $m<\infty$. Thus $\limsup_{n\to\infty}\mathrm{P}^*(X_n\notin K^\delta)\le\limsup_{n\to\infty}\mathrm{P}^*(X_n\notin\cap_{i=1}^m K_i)\le\zeta+\sum_{i=1}^m\eta_m\le 2\zeta$. Since this result holds for all $\delta>0$, we now have that $\liminf_{n\to\infty}\mathrm{P}^*(X_n\in K^\delta)\ge 1-2\zeta$ for all $\delta>0$. Asymptotic tightness follows since $\zeta$ is arbitrary.□

**Proof of Lemma 7.20.** The fact that (i) implies (ii) we leave as an exercise. Assume (ii). Then there exists a sequence of finite partitions $\mathcal{T}_1,\mathcal{T}_2,\ldots$ such that

$$\limsup_{n\to\infty}\mathrm{P}^*\left(\sup_{U\in\mathcal{T}_k}\sup_{s,t\in U}|X_n(s)-X_n(t)|>2^{-k}\right)<2^{-k},$$

for all integers $k\ge 1$. Here, each partition $\mathcal{T}_k$ is a collection of disjoint sets $U_1,\ldots,U_{m_k}$ with $T=\cup_{i=1}^{m_k}U_i$. Without loss of generality, we can insist that the $\mathcal{T}_k$ are nested in the sense that any $U\in\mathcal{T}_k$ is a union of sets in $\mathcal{T}_{k+1}$. Such a nested sequence of partitions is easy to construct from any other sequence of partitions $\mathcal{T}_k^*$ by letting $\mathcal{T}_k$ consist of all nontrivial intersections of all sets in $\mathcal{T}_1^*$ up through and including $\mathcal{T}_k^*$. Let $\mathcal{T}_0$ denote the partition consisting of the single set $T$.

For any $s,t\in T$, define $K(s,t)\equiv\sup\{k:s,t\in U\text{ for some }U\in\mathcal{T}_k\}$ and $\rho(s,t)\equiv 2^{-K(s,t)}$. Also, for any $\delta>0$, let $J(\delta)\equiv\inf\{k:2^{-k}<\delta\}$. It is not hard to verify that for any $s,t\in T$ and $\delta>0$, $\rho(s,t)<\delta$ if and only if $s,t\in U$ for some $U\in\mathcal{T}_{J(\delta)}$. Thus

$$\sup_{s,t\in T:\rho(s,t)<\delta} |X_n(s) - X_n(t)| = \sup_{U\in\mathcal{T}_{J(\delta)}} \sup_{s,t\in U} |X_n(s) - X_n(t)|$$

for all $0 < \delta \leq 1$. Since $J(\delta) \to \infty$ as $\delta \downarrow 0$, we now have that $X_n$ is asymptotically $\rho$-equicontinuous in probability, as long as $\rho$ is a pseudometric. The only difficulty here is to verify that the triangle inequality holds for $\rho$, which we leave as an exercise. It is easy to see that $T$ is totally bounded with respect to $\rho$, and (i) follows.□

**Proof of Lemma 7.21.** We first prove (iii). Assume that $X_n \overset{\text{as*}}{\to} X$, and define $A_n^k \equiv \{\sup_{m\geq n} d(X_m, X)^* > 1/k\}$. For each integer $k \geq 1$, we have $P(A_n^k) \downarrow 0$ as $n \to \infty$. Now fix $\epsilon > 0$, and note that for each $k \geq 1$ we can choose an $n_k$ so that $P(A_n^k) \leq \epsilon/2^k$. Let $A = \Omega - \cup_{k=1}^{\infty} A_{n_k}^k$, and observe that, by this construction, $P(A) \geq 1 - \epsilon$ and $d(X_n, X)^* \leq 1/k$, for all $n \geq n_k$ and all $\omega \in A$. Thus $X_n$ converges to $X$ almost uniformly since $\epsilon$ is arbitrary. Assume now that $X_n$ converges to $X$ almost uniformly. Fix $\epsilon > 0$, and let $A$ be measurable with $P(A) \geq 1 - \epsilon$ and $d(X_n, X) \to 0$ uniformly over $\omega \in A$. Fix $\eta > 0$, and note that $\eta \geq (d(X_n, X)1\{A\})^*$ for all sufficiently large $n$, since $\eta$ is measurable and satisfies $\eta \geq d(X_n, X)1\{A\}$ for sufficiently large $n$. Now let $S, T : \Omega \mapsto [0, \infty)$ be maps with $S$ measurable and $T^*$ bounded. The for any $c > 0$, $[(S + c)T]^* \leq (S + c)T^*$, and $(S + c)T^* \leq [(S + c)T]^*$ since $T^* \leq [(S + c)T]^*/(S + c)$. Hence $[(S + c)T]^* = (S + c)T^*$. By letting $c \downarrow 0$, we obtain that $(ST)^* = ST^*$. Hence $d(X_n, X)^*1\{A\} = (d(X_n, X)1\{A\})^* \leq \eta$ for all $n$ large enough, and thus $d(X_n, X)^* \to 0$ for almost all $\omega \in A$. Since $\epsilon$ is arbitrary, $X_n \overset{\text{as*}}{\to} X$.

Now assume that $X_n \overset{\text{as*}}{\to} X$. This clearly implies that $\sup_{m\geq n} d(X_m, X) \overset{\text{P}}{\to} 0$. Fix $\epsilon > 0$. Generate a subsequence $n_k$, by finding, for each integer $k \geq 1$, an integer $n_k \geq n_{k-1} \geq 1$ which satisfies $\text{P}^*(\sup_{m\geq n_k} d(X_m, X) > 1/k) \leq \epsilon/2^k$. Call the set inside this outer probability statement $A_k$, and define $A = \Omega - \cap_{k=1}^{\infty} A_k^*$. Now $\text{P}(A) \geq 1 - \epsilon$, and for each $\omega \in A$ and all $m \geq n_k$, $d(X_m, X) \leq 1/k$ for all $k \geq 1$. Hence $\sup_{\omega\in A} d(X_n, X)(\omega) \to 0$, as $n \to \infty$. Thus $X_n$ converges almost uniformly to $X$, since $\epsilon$ is arbitrary, and therefore $X_n \overset{\text{as*}}{\to} X$. Thus we have proven (iii).

We now prove (i). It is easy to see that $X_n$ converging to $X$ almost uniformly will imply $X_n \overset{\text{P}}{\to} X$. Thus (i) follows from (iii). We next prove (ii). Assume $X_n \overset{\text{P}}{\to} X$. Construct a subsequence $1 \leq n_1 < n_2 < \cdots$ so that $\text{P}(d(X_{n_j}, X)^* > 1/j) < 2^{-j}$ for all integers $j \geq 1$. Then

$$\text{P}(d(X_{n_j}, X)^* > 1/j, \text{ for infinitely many } j) = 0$$

by the Borel-Cantelli lemma. Hence $X_{n_j} \overset{\text{as*}}{\to} X$ as a sequence in $j$. This now implies that every sequence has an outer almost surely convergent subsequence. Assume now that every subsequence has an outer almost surely convergent subsequence. By (i), the almost surely convergent subsequences also converge in probability. Hence $X_n \overset{\text{P}}{\to} X$.□

## 7.5 Exercises

**7.5.1.** Show that $\mathcal{F}$ defined in the proof of Lemma 7.3 is a vector lattice, an algebra, and separates points in $\mathbb{D}$.

**7.5.2.** Show that the set $K$ defined in the proof of Lemma 7.10 in Section 7.2 is compact.

**7.5.3.** In the setting of the proof of Lemma 7.14, show that $\pi_1 K$ and $\pi_2 K$ are both compact whenever $K \in \mathbb{D} \times \mathbb{E}$ is compact.

**7.5.4.** In the proof of Lemma 7.18 (given in Section 7.4), show that the set $K = \cap_{m=1}^{\infty} K_m$ is closed.

**7.5.5.** Suppose that $T_n$ and $T_0$ are subsets of a semimetric space $(T, \rho)$ such that $T_n \to T_0$ in the sense that

(i) Every $t \in T_0$ is the limit of a sequence $t_n \in T_n$;

(ii) For every closed $S \subset T - T_0$, $S \cap T_n = \emptyset$ for all $n$ large enough.

Suppose that $X_n$ and $X$ are stochastic processes indexed by $T$ for which $X_n \rightsquigarrow X$ in $\ell^\infty(T)$ and $X$ is Borel measurable with $\mathrm{P}(X \in UC(T, \rho)) = 1$, where $(T, \rho)$ is not necessarily totally bounded. Show that $\sup_{t \in T_n} X_n(t) \rightsquigarrow \sup_{t \in T_0} X(t)$. Hint: show first that for any $x_n \to x$, where $\{x_n\} \in \ell^\infty(T)$ and $x \in UC(T, \rho)$, we have $\lim_{n \to \infty} \sup_{t \in T_n} x_n(t) = \sup_{t \in T_0} x(t)$.

**7.5.6.** Complete the proof of Lemma 7.20:

1. Show that (i) implies (ii).
2. Show that for any $s, t \in T$ and $\delta > 0$, $\rho(s,t) < \delta$ if and only if $s, t \in U$ for some $U \in \mathcal{T}_{J(\delta)}$, where $\rho(s,t) \equiv 2^{-K(s,t)}$ and $K$ is as defined in the proof.
3. Verify that the triangle inequality holds for $\rho$.

**7.5.7.** Let $X_n$ and $X$ be maps into $\mathbb{R}$ with $X$ Borel measurable. Show the following:

(i) $X_n \overset{\text{as*}}{\to} X$ if and only if both $X_n^*$ and $X_{n*} \equiv (X_n)_*$ converge almost surely to $X$.

(ii) $X_n \rightsquigarrow X$ if and only if $X_n^* \rightsquigarrow X$ and $X_{n*} \rightsquigarrow X$.

Hints: For (i), first show $|X_n - X|^* = |X_n^* - X| \vee |X_{n*} - X|$. For (ii), assume $X_n \rightsquigarrow X$ and apply the extended almost sure representation theorem to find a new probability space and a perfect sequence of maps $\phi_n$ such that $\tilde{X}_n = X_n \circ \phi_n \overset{\text{as*}}{\to} \tilde{X}$. By (i), $\tilde{X}_n^* \to \tilde{X}$ almost surely, and thus $\tilde{X}_n^* \rightsquigarrow \tilde{X}$. Since $\phi_n$ is perfect, $\tilde{X}_n^* = X_n^* \circ \phi_n$; and thus $\mathrm{E} f(\tilde{X}_n^*) = \mathrm{E} f(X_n^*)$ for every

measurable $f$. Hence $X_n^* \rightsquigarrow X$. Now show $X_{n*} \rightsquigarrow X$. For the converse, use the facts that $\mathrm{P}(X_n^* \le x) \le \mathrm{P}^*(X_n \le x) \le \mathrm{P}(X_{n*} \le x)$ and that distributions for real random variable are completely determined by their cumulative distribution functions.

**7.5.8.** Using the ideas in the proof of Proposition 7.27, prove Lemma 4.2.

## 7.6 Notes

Many of the ideas and results of this chapter come from Chapters 1.3–1.5 and 1.9–1.12 of VW specialized to sequences (rather than nets). Lemma 7.1 is a composite of Lemmas 1.3.12 and Theorem 1.12.2 of VW, while Lemma 7.4 is Lemma 1.5.3 of VW. Components (i)–(vii) of the portmanteau theorem are a specialization to sequences of the portmanteau theorem in Chapter 1.3 of VW. Theorem 7.7, Lemmas 7.8, 7.9, and 7.12, and Theorem 7.13 correspond to VW Theorems 1.3.6 and 1.3.10, Lemmas 1.3.13 and 1.3.8, and Theorem 1.3.9, respectively. Lemma 7.14 is a composite of Lemmas 1.4.3 and 1.4.4 of VW. Lemma 7.16 and Theorem 7.17 are essentially VW Lemma 1.5.2 and Theorem 1.5.4. Lemmas 7.21 and 7.23 and Theorem 7.24 are specializations to sequences of VW Lemmas 1.9.2 and 1.10.2 and Theorem 1.11.1. Theorem 7.26 is a composition of VW Theorem 1.10.4 and Addendum 1.10.5 applied to sequences, while Lemma 7.28 is essentially Proposition 1.10.12 of VW.

Proposition 7.5 on Gaussian processes is essentially a variation of Lemma 3.9.8 of VW. Proposition 7.27 is a modification of Lemma A.3 of Bilias, Gu and Ying (1997) who use this result to obtain weak convergence of the proportional hazards regression parameter in a continuously monitored sequential clinical trial.

# 8
# Empirical Process Methods

Recall from Section 2.2.2 the concepts of bracketing and uniform entropy along with the corresponding Glivenko-Cantelli and Donsker theorems. We now briefly review the set-up. Given a probability space $(\Omega, \mathcal{A}, P)$, the data of interest consist of $n$ independent copies $X_1, \ldots, X_n$ of a map $X : \Omega \mapsto \mathcal{X}$, where $\mathcal{X}$ is the sample space. We are interested in studying the limiting behavior of empirical processes indexed by classes $\mathcal{F}$ of functions $f : \mathcal{X} \mapsto \mathbb{R}$ which are measurable in the sense that each composite map $f(X) : \Omega \mapsto \mathcal{X} \mapsto \mathbb{R}$ is $\mathcal{A}$-measurable. With $\mathbb{P}_n$ denoting the empirical measure based on the data $X_1, \ldots, X_n$, the empirical process of interest is $\mathbb{P}_n f$ viewed as a stochastic process indexed by $f \in \mathcal{F}$.

A class $\mathcal{F}$ is $P$-Glivenko-Cantelli if $\|\mathbb{P}_n - P\|_{\mathcal{F}} \overset{\text{as*}}{\to} 0$, where for any $u \in \ell^\infty(T)$, $\|u\|_T \equiv \sup_{t \in T} |u(t)|$. We say $\mathcal{F}$ is *weak $P$-Glivenko-Cantelli*, if the outer almost sure convergence is replaced by convergence in probability. Sometimes, for clarification, we will call a $P$-Glivenko-Cantelli class a *strong $P$-Glivenko-Cantelli* class to remind ourselves that the convergence is outer almost sure. A class $\mathcal{F}$ is $P$-Donsker if $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ weakly in $\ell^\infty(\mathcal{F})$, where $\mathbb{G}$ is a tight Brownian bridge. Of course, the $P$ prefix can be dropped if the context is clear.

As mentioned in Section 4.2.1, these ideas also apply directly to i.i.d. samples of stochastic processes. In this setting, $X$ has the form $\{X(t), t \in T\}$, where $X(t)$ is measurable for each $t \in T$, and $\mathcal{X}$ is typically $\ell^\infty(T)$. We say that $X$ is *$P$-Glivenko-Cantelli* if $\sup_{t \in T} |(\mathbb{P}_n - P)X(t)| \overset{\text{as*}}{\to} 0$ and that $X$ is *$P$-Donsker* if $\mathbb{G}_n X$ converges weakly to a tight Gaussian process. This is exactly equivalent to considering whether the class $\mathcal{F}_T \equiv \{f_t, t \in T\}$,

where $f_t(x) \equiv x(t)$ for all $x \in \mathcal{X}$ and $t \in T$, is Glivenko-Cantelli or Donsker. The limiting Brownian bridge $\mathbb{G}$ for the class $\mathcal{F}_T$ has covariance $P[(\mathbb{G}f_s)(\mathbb{G}f_t)] = P[(X(s) - PX(s))(X(t) - PX(t))]$. This duality between the function class and stochastic process viewpoint will prove useful from time to time, and which approach we take will depend on the setting.

The main goal of this chapter is to present the empirical process techniques needed to prove the Glivenko-Cantelli and Donsker theorems of Section 2.2.2. The approach we take is guided by Chapters 2.2–2.5 of VW, although we leave out many technical details. The most difficult step in these proofs is going from point-wise convergence to uniform convergence. Maximal inequalities are very useful tools for accomplishing this step. For uniform entropy results, an additional tool, symmetrization, is also needed. To use symmetrization, several measurability conditions are required on the class of function $\mathcal{F}$ beyond the usual requirement that each $f \in \mathcal{F}$ be measurable. In the sections presenting the Glivenko-Cantelli and Donsker theorem proofs, results for bracketing entropy are presented before the uniform entropy results.

## 8.1 Maximal Inequalities

We first present several results about Orlicz norms which are useful for controlling the size of the maximum of a finite collection of random variables. Several maximal inequalities for stochastic processes will be given next. These inequalities include a general maximal inequality for separable stochastic processes and a maximal inequality for sub-Gaussian processes. The results will utilize Orlicz norms combined with a method known as *chaining*. The results of this section will play a key role in the proofs of the Donsker theorems developed later on in this chapter.

### *8.1.1 Orlicz Norms and Maxima*

A very useful class of norms for random variables used in maximal inequalities are the *Orlicz norms*. For a nondecreasing, nonzero convex function $\psi : [0, \infty] \mapsto [0, \infty]$, with $\psi(0) = 0$, the Orlicz norm $\|X\|_\psi$ of a real random variable $X$ is defined as

$$\|X\|_\psi \equiv \inf \left\{ c > 0 : \mathrm{E}\psi\left(\frac{|X|}{c}\right) \leq 1 \right\},$$

where the norm takes the value $\infty$ if no finite $c$ exists for which $\mathrm{E}\psi(|X|/c) \leq 1$. Exercise 8.5.1 below verifies that $\|\cdot\|_\psi$ is indeed a norm on the space of random variables with $\|X\|_\psi < \infty$. The Orlicz norm $\|\cdot\|_\psi$ is also called the *$\psi$-norm*, in order to specify the choice of $\psi$. When $\psi$ is of the form $x \mapsto x^p$, where $p \geq 1$, the corresponding Orlicz norm is just the $L_p$-norm

$\|X\|_p \equiv (\mathrm{E}|X|^p)^{1/p}$. For maximal inequalities, Orlicz norms defined with $\psi_p(x) \equiv e^{x^p} - 1$, for $p \geq 1$, are of greater interest because of their sensitivity to behavior in the tails. Clearly, since $x^p \leq \psi_p(x)$, we have $\|X\|_p \leq \|X\|_{\psi_p}$. Also, by the series representation for exponentiation, $\|X\|_p \leq p!\|X\|_{\psi_1}$ for all $p \geq 1$. The following result shows how Orlicz norms based on $\psi_p$ relate fairly precisely to the tail probabilities:

LEMMA 8.1 *For a real random variable $X$ and any $p \in [1, \infty)$, the following are equivalent:*

*(i)* $\|X\|_{\psi_p} < \infty$.

*(ii) There exist constants $0 < C, K < \infty$ such that*

$$\mathrm{P}(|X| > x) \quad \leq \quad Ke^{-Cx^p}, \quad \textit{for all } x > 0. \tag{8.1}$$

*Moreover, if either condition holds, then $K = 2$ and $C = \|X\|_{\psi_p}^{-p}$ satisfies (8.1), and, for any $C, K \in (0, \infty)$ satisfying (8.1), $\|X\|_{\psi_p} \leq ((1 + K)/C)^{1/p}$.*

**Proof.** Assume (i). Then $\mathrm{P}(|X| > x)$ equals

$$\mathrm{P}\left\{\psi_p(|X|/\|X\|_{\psi_p}) \geq \psi_p(x/\|X\|_{\psi_p})\right\} \quad \leq \quad 1 \wedge \left(\frac{1}{\psi_p(x/\|X\|_{\psi_p})}\right), \tag{8.2}$$

by Markov's inequality. By Exercise 8.5.2 below, $1 \wedge (e^u - 1)^{-1} \leq 2e^{-u}$ for all $u > 0$. Thus the right-hand-side of (8.2) is bounded above by (8.1) with $K = 2$ and $C = \|X\|_{\psi_p}^{-p}$. Hence (ii) and the first half of the last sentence of the lemma follow. Now assume (ii). For any $c \in (0, C)$, Fubini's theorem gives us

$$\begin{aligned}
\mathrm{E}\left(e^{c|X|^p} - 1\right) &= \mathrm{E}\int_0^{|X|^p} ce^{cs} ds \\
&= \int_0^\infty \mathrm{P}(|X| > s^{1/p}) ce^{cs} ds \\
&\leq \int_0^\infty Ke^{-Cs} ce^{cs} ds \\
&= Kc/(C - c),
\end{aligned}$$

where the inequality follows from the assumption. Now $Kc/(C - c) \leq 1$ whenever $c \leq C/(1 + K)$ or, in other words, whenever $c^{-1/p} \geq ((1 + K)/C)^{1/p}$. This implies (i) and the rest of the lemma.□

An important use for Orlicz norms is to control the behavior of maxima. This control is somewhat of an extension of the following simple result for $L_p$-norms: For any random variables $X_1, \ldots, X_m$, $\|\max_{1 \leq i \leq m} X_i\|_p$

$$\leq \left( \mathrm{E} \max_{1\leq i\leq m} |X_i|^p \right)^{1/p} \leq \left( \mathrm{E} \sum_{i=1}^{m} |X_i|^p \right)^{1/p} \leq m^{1/p} \max_{1\leq i\leq m} \|X_i\|_p.$$

The following lemma shows that a similar result holds for certain Orlicz norms but with the $m^{1/p}$ replaced with a constant times $\psi^{-1}(m)$:

LEMMA 8.2 *Let $\psi : [0,\infty) \mapsto [0,\infty)$ be convex, nondecreasing and nonzero, with $\psi(0) = 0$ and $\limsup_{x,y\to\infty} \psi(x)\psi(y)/\psi(cxy) < \infty$ for some constant $c < \infty$. Then, for any random variables $X_1, \ldots, X_m$,*

$$\left\| \max_{1\leq i\leq m} X_i \right\|_\psi \leq K\psi^{-1}(m) \max_{1\leq i\leq m} \|X_i\|_\psi,$$

*where the constant $K$ depends only on $\psi$.*

**Proof.** We first make the stronger assumption that $\psi(1) \leq 1/2$ and that $\psi(x)\psi(y) \leq \psi(cxy)$ for all $x, y \geq 1$. Under this stronger assumption, we also have $\psi(x/y) \leq \psi(cx)/\psi(y)$ for all $x \geq y \geq 1$. Hence, for any $y \geq 1$ and $k > 0$,

$$\begin{aligned} \max_{1\leq i\leq m} \psi\left(\frac{|X_i|}{ky}\right) &\leq \max_i \left[ \frac{\psi(c|X_i|/k)}{\psi(y)} 1\left\{\frac{|X_i|}{ky} \geq 1\right\} \right. \\ &\qquad \left. + \psi\left(\frac{|X_i|}{ky}\right) 1\left\{\frac{|X_i|}{ky} < 1\right\} \right] \\ &\leq \sum_{i=1}^{m} \left[ \frac{\psi(c|X_i|/k)}{\psi(y)} \right] + \psi(1). \end{aligned}$$

In the summation, set $k = c \max_i \|X_i\|_\psi$ and take expectations of both sides to obtain

$$\mathrm{E}\psi\left(\frac{\max_i |X_i|}{ky}\right) \leq \frac{m}{\psi(y)} + \frac{1}{2}.$$

With $y = \psi^{-1}(2m)$, the right-hand-side is $\leq 1$. Thus $\|\max_i |X_i|\|_\psi \leq c\psi^{-1}(2m) \max_i \|X_i\|_\psi$. Since $\psi$ is convex and $\psi(0) = 0$, $x \mapsto \psi^{-1}(x)$ is concave and one-to-one for $x > 0$. Thus $\psi^{-1}(2m) \leq 2\psi^{-1}(m)$, and the result follows with $K = 2c$ for the special $\psi$ functions specified at the beginning of the proof.

By Exercise 8.5.3 below, we have for any $\psi$ satisfying the conditions of the lemma, that there exists constants $0 < \sigma \leq 1$ and $\tau > 0$ such that $\phi(x) \equiv \sigma\psi(\tau x)$ satisfies $\phi(1) \leq 1/2$ and $\phi(x)\phi(y) \leq \phi(cxy)$ for all $x, y \geq 1$. Furthermore, for this $\phi$, $\phi^{-1}(u) \leq \psi^{-1}(u)/(\sigma\tau)$, for all $u > 0$, and, for any random variable $X$, $\|X\|_\psi \leq \|X\|_\phi/(\sigma\tau) \leq \|X\|_\psi/\sigma$. Hence

$$\begin{aligned} \sigma\tau \left\|\max_i X_i\right\|_\psi &\leq \left\|\max_i X_i\right\|_\phi \\ &\leq 2c\phi^{-1}(m) \max_i \|X_i\|_\phi \leq \frac{2c}{\sigma}\psi^{-1}(m) \max_i \|X_i\|_\psi, \end{aligned}$$

and the desired result follows with $K = 2c/(\sigma^2\tau)$.□

An important consequence of Lemma 8.2 is that maximums of random variables with bounded $\psi$-norm grow at the rate $\psi^{-1}(m)$. Based on Exercise 8.5.4, $\psi_p$ satisfies the conditions of Lemma 8.2 with $c = 1$, for any $p \in [1, \infty)$. The implication in this situation is that the growth of maxima is at most logarithmic, since $\psi_p^{-1}(m) = (\log(m+1))^{1/p}$. These results will prove quite useful in the next section.

We now present an inequality for collections $X_1, \ldots, X_m$ of random variables which satisfy

$$P(|X_i| > x) \quad \leq \quad 2e^{-\frac{1}{2}\frac{x^2}{b+ax}}, \quad \text{for all } x > 0, \tag{8.3}$$

for $i = 1, \ldots, m$ and some $a, b \geq 0$. This setting will arise later in the development of a Donsker theorem based on bracketing entropy.

LEMMA 8.3 *Let $X_1, \ldots, X_m$ be random variables that satisfy the tail bound (8.3) for $1 \leq i \leq m$ and some $a, b \geq 0$. Then*

$$\left\| \max_{1 \leq i \leq m} |X_i| \right\|_{\psi_1} \leq K \left\{ a \log(1+m) + \sqrt{b}\sqrt{\log(1+m)} \right\},$$

*where the constant $K$ is universal, in the sense that it does not depend on $a$, $b$, or on the random variables.*

**Proof.** Assume for now that $a, b > 0$. The condition implies for all $x \leq b/a$ the upper bound $2\exp(-x^2/(4b))$ for $P(|X_i| > x)$, since in this case $b + ax \leq 2b$. For all $x > b/a$, the condition implies an upper bound of $2\exp(-x/(4a))$, since $b/x + a \leq 2a$ in this case. This implies that $P(|X_i|1\{|X_i| \leq b/a\} > x) \leq 2\exp(-x^2/(4b))$ and $P(|X_i|1\{|X_i| > b/a\} > x) \leq 2\exp(-x/(4a))$ for all $x > 0$. Hence, by Lemma 8.1, the Orlicz norms $\| |X_i|1\{|X_i| \leq b/a\}\|_{\psi_2}$ and $\| |X_i|1\{|X_i| > b/a\}\|_{\psi_1}$ are bounded by $\sqrt{12b}$ and $12a$, respectively. The result now follows by Lemma 8.2 combined with the inequality

$$\|\max_i |X_i| \,\|_{\psi_1} \leq \| \max_i [|X_i|1\{|X_i| \leq b/a\}] \,\|_{\psi_1} + \| \max_i [|X_i|1\{|X_i| > b/a\}] \,\|_{\psi_2},$$

where the replacement of $\psi_1$ with $\psi_2$ in the last term follows since $\psi_p$ norms increase in $p$.

Suppose now that $a > 0$ but $b = 0$. Then the tail bound (8.3) holds for all $b > 0$, and the result of the lemma is thus true for all $b > 0$. The desired result now follows by letting $b \downarrow 0$. A similar argument will verify that the result holds when $a = 0$ and $b > 0$. Finally, the result is trivially true when $a = b = 0$ since, in this case, $X_i = 0$ almost surely for $i = 1, \ldots, m$.□

### *8.1.2 Maximal Inequalities for Processes*

The goals of this section are to first establish a general maximal inequality for *separable* stochastic processes and then specialize the result to *sub-Gaussian* processes. A stochastic process $\{X(t), t \in T\}$ is separable when

there exists a countable subset $T_* \subset T$ such that $\sup_{t\in T} \inf_{s\in T_*} |X(t) - X(s)| = 0$ almost surely. For example, any cadlag process indexed by a closed interval in $\mathbb{R}$ is separable because the rationals are a separable subset of $\mathbb{R}$. The need for separability of certain processes in the Glivenko-Cantelli and Donsker theorems is hidden in other conditions of the involved theorems, and direct verification of separability is seldom required in statistical applications.

A stochastic process is sub-Gaussian when

$$(8.4)\quad P(|X(t) - X(s)| > x) \quad \leq \quad 2e^{-\frac{1}{2}x^2/d^2(s,t)}, \quad \text{for all } s, t \in T,\ x > 0,$$

for a semimetric $d$ on $T$. In this case, we say that $X$ is sub-Gaussian with respect to $d$. An important example of a separable sub-Gaussian stochastic process, the Rademacher process, will be presented at the end of this section. These processes will be utilized later in this chapter in the development of a Donsker theorem based on uniform entropy. Another example of a sub-Gaussian process is Brownian motion on $[0, 1]$, which can easily be shown to be sub-Gaussian with respect to $d(s, t) = |s - t|^{1/2}$. Because the sample paths are continuous, Brownian motion is also separable.

The conclusion of Lemma 8.2 above is not immediately useful for maximizing $X(t)$ over $t \in T$ since a potentially infinite number of random variables is involved. However, a method called *chaining*, which involves linking up increasingly refined finite subsets of $T$ and repeatedly applying Lemma 8.2, does make such maximization possible in some settings. The technique depends on the *metric entropy* of the index set $T$ based on the semimetric $d(s, t) = \|X(s) - X(t)\|_\psi$.

For an arbitrary semimetric space $(T, d)$, the *covering number* $N(\epsilon, T, d)$ is the minimal number of closed $d$-balls of radius $\epsilon$ required to cover $T$. The *packing number* $D(\epsilon, T, d)$ is the maximal number of points that can fit in $T$ while maintaining a distance greater than $\epsilon$ between all points. When the choice of index set $T$ is clear by context, the notation for covering and packing numbers will be abbreviated as $N(\epsilon, d)$ and $D(\epsilon, d)$, respectively. The associated *entropy numbers* are the respective logarithms of the covering and packing numbers. Taken together, these concepts define metric entropy.

For a semimetric space $(T, d)$ and each $\epsilon > 0$,

$$N(\epsilon, d) \leq D(\epsilon, d) \leq N(\epsilon/2, d).$$

To see this, note that there exists a minimal subset $T_\epsilon \subset T$ such that the cardinality of $T_\epsilon = D(\epsilon, d)$ and the minimum distance between distinct points in $T_\epsilon$ is $> \epsilon$. If we now place closed $\epsilon$-balls around each point in $T_\epsilon$, we have a covering of $T$. If this were not true, there would exist a point $t \in T$ which has distance $> \epsilon$ from all the points in $T_\epsilon$, but this would mean that $D(\epsilon, d) + 1$ points can fit into $T$ while still maintaining a separation $> \epsilon$ between all points. But this contradicts the maximality of $D(\epsilon, d)$. Thus

$N(\epsilon, d) \leq D(\epsilon, d)$. Now note that no ball of radius $\leq \epsilon/2$ can cover more than one point in $T_\epsilon$, and thus at least $D(\epsilon, d)$ closed $\epsilon/2$-balls are needed to cover $T_\epsilon$. Hence $D(\epsilon, d) \leq N(\epsilon/2, d)$.

This forgoing discussion reveals that covering and packing numbers are essentially equivalent in behavior as $\epsilon \downarrow 0$. However, it turns out to be slightly more convenient for our purposes to focus on packing numbers in this section. Note that $T$ is totally bounded if and only if $D(\epsilon, d)$ is finite for each $\epsilon > 0$. The success of the following maximal inequality depends on how fast $D(\epsilon, d)$ increases as $\epsilon \downarrow 0$:

THEOREM 8.4 *(General maximal inequality) Let $\psi$ satisfy the conditions of Lemma 8.2, and let $\{X(t), t \in T\}$ be a separable stochastic process with $\|X(s) - X(t)\|_\psi \leq rd(s,t)$, for all $s, t \in T$, some semimetric $d$ on $T$, and a constant $r < \infty$. Then for any $\eta, \delta > 0$,*

$$\left\| \sup_{s,t \in T : d(s,t) \leq \delta} |X(s) - X(t)| \right\|_\psi \leq K \left[ \int_0^\eta \psi^{-1}(D(\epsilon, d)) d\epsilon + \delta \psi^{-1}(D^2(\eta, d)) \right],$$

*for a constant $K < \infty$ which depends only on $\psi$ and $r$. Moreover,*

$$\left\| \sup_{s,t \in T} |X(s) - X(t)| \right\|_\psi \leq 2K \int_0^{\operatorname{diam} T} \psi^{-1}(D(\epsilon, d)) d\epsilon,$$

*where $\operatorname{diam} T \equiv \sup_{s,t \in T} d(s,t)$ is the diameter of $T$.*

Before we present the proof of this theorem, recall in the discussion following the proof of Lemma 8.2 that $\psi_p$-norms, for any $p \in [1, \infty)$, satisfy the conditions of this lemma. The case $p = 2$, which applies to sub-Gaussian processes, is of most interest to us and is explicitly evaluated in Corollary 8.5 below. This corollary plays a key role in the proof of the Donsker theorem for uniform entropy (see Theorem 8.19 in Section 8.4).

**Proof of Theorem 8.4.** Note that if the first integral were infinite, the inequalities would be trivially true. Hence we can, without loss of generality assume that the packing numbers and associated integral are bounded. Construct a sequence of finite nested sets $T_0 \subset T_1 \subset \cdots \subset T$ such that for each $T_j$, $d(s,t) > \eta 2^{-j}$ for every distinct $s, t \in T_j$, and that each $T_j$ is "maximal" in the sense that no additional points can be added to $T_j$ without violating the inequality. Note that by the definition of packing numbers, the number of points in $T_j$ is bounded above by $D(\eta 2^{-j}, d)$.

Now we will do the chaining part of the proof. Begin by "linking" every point $t_{j+1} \in T_{j+1}$ to one and only one $t_j \in T_j$ such that $d(t_j, t_{j+1}) \leq \eta 2^{-j}$, for all points in $T_{j+1}$. Continue this process to link all points in $T_j$ with points in $T_{j-1}$, and so on, to obtain for every $t_{j+1}$ ($\in T_{j+1}$) a chain $t_{j+1}, t_j, t_{j-1}, \ldots, t_0$ that connects to a point in $T_0$. For any integer $k \geq 0$ and arbitrary points $s_{k+1}, t_{k+1} \in T_{k+1}$, the difference in increments along their respective chains connecting to $s_0, t_0$ can be bounded as follows:

$$
\begin{aligned}
&|\{X(s_{k+1}) - X(t_{k+1})\} - \{X(s_0) - X(t_0)\}| \\
&\qquad = \left| \sum_{j=0}^{k} \{X(s_{j+1}) - X(s_j)\} - \sum_{j=0}^{k} \{X(t_{j+1}) - X(t_j)\} \right| \\
&\qquad \leq 2 \sum_{j=0}^{k} \max |X(u) - X(v)|,
\end{aligned}
$$

where for fixed $j$ the maximum is taken over all links $(u, v)$ from $T_{j+1}$ to $T_j$. Hence the $j$th maximum is taken over at most the cardinality of $T_{j+1}$ links, with each link having $\|X(u) - X(v)\|_\psi$ bounded by $rd(u, v) \leq r\eta 2^{-j}$. By Lemma 8.2, we have for a constant $K_0 < \infty$ depending only on $\psi$ and $r$,

$$
\begin{aligned}
(8.5) \quad &\left\| \max_{s,t \in T_{k+1}} |\{X(s) - X(s_0)\} - \{X(t) - X(t_0)\}| \right\|_\psi \\
&\qquad \leq K_0 \sum_{j=0}^{k} \psi^{-1}(D(\eta 2^{-j-1}, d))\eta 2^{-j} \\
&\qquad = 4K_0 \sum_{j=0}^{k} \psi^{-1}(D(\eta 2^{-k+j-1}, d))\eta 2^{-k+j-2} \\
&\qquad \leq 4\eta K_0 \int_0^1 \psi^{-1}(D(\eta u, d))du \\
&\qquad = 4K_0 \int_0^\eta \psi^{-1}(D(\epsilon, d))d\epsilon.
\end{aligned}
$$

In this bound, $s_0$ and $t_0$ depend on $s$ and $t$ in that they are the endpoints of the chains starting at $s$ and $t$, respectively.

The maximum of the increments $|X(s_{k+1}) - X(t_{k+1})|$, over all $s_{k+1}$ and $t_{k+1}$ in $T_{k+1}$ with $d(s_{k+1}, t_{k+1}) < \delta$, is bounded by the left-hand-side of (8.5) plus the maximum of the discrepancies at the ends of the chains $|X(s_0) - X(t_0)|$ for those points in $T_{k+1}$ which are less than $\delta$ apart. For every such pair of endpoints $s_0, t_0$ of chains starting at two points in $T_{k+1}$ within distance $\delta$ of each other, choose one and only 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 results in at most $D^2(\eta, d)$ pairs. Now,

$$
\begin{aligned}
(8.6) \quad |X(s_0) - X(t_0)| \leq\ & |\{X(s_0) - X(s_{k+1})\} - \{X(t_0) - X(t_{k+1})\}| \\
& + |X(s_{k+1}) - X(t_{k+1})|.
\end{aligned}
$$

Take the maximum of (8.6) over all pairs of endpoints $s_0, t_0$. The maximum of the first term of the right-hand-side of (8.6) is bounded by the left-hand-side of (8.5). The maximum of the second term of the right-hand-side of (8.6) is the maximum of $D^2(\eta, d)$ terms with $\psi$-norm bounded by

$r\delta$. By Lemma 8.2, this maximum is bounded by some constant $C$ times $\delta\psi^{-1}(D^2(\eta,d))$. Combining this with (8.5), we obtain

$$\begin{aligned}\left\| \max_{s,t\in T_{k+1}:d(s,t)<\delta} |X(s)-X(t)| \right\|_\psi \\ \leq \quad 8K_0 \int_0^\eta \psi^{-1}(D(\epsilon,d))d\epsilon + C\delta\psi^{-1}(D^2(\eta,d)).\end{aligned}$$

By the fact that the right-hand-side does not depend on $k$, we can replace $T_{k+1}$ with $T_\infty = \cup_{j=0}^\infty T_j$ by the monotone convergence theorem. If we can verify that taking the supremum over $T_\infty$ is equivalent to taking the supremum over $T$, then the first conclusion of the theorem follows with $K = (8K_0) \vee C$.

Since $X$ is separable, there exists a countable subset $T_* \subset T$ such that $\sup_{t\in T} \inf_{s\in T_*} |X(t)-X(s)| = 0$ almost surely. Let $\Omega_*$ denote the subset of the sample space of $X$ for which this supremum is zero. Accordingly $\mathrm{P}(\Omega_*) = 1$. Now, for any point $t$ and sequence $\{t_n\}$ in $T$, it is easy to see that $d(t,t_n) \to 0$ implies $|X(t)-X(t_n)| \to 0$ almost surely (see Exercise 8.5.5 below). For each $t \in T_*$, let $\Omega_t$ be the subset of the sample space of $X$ for which $\inf_{s\in T_\infty} |X(s)-X(t)| = 0$. Since $T_\infty$ is a dense subset of the semimetric space $(T,d)$, $\mathrm{P}(\Omega_t) = 1$. Letting $\tilde{\Omega} \equiv \Omega_* \cap (\cap_{t\in T_*}\Omega_t)$, we now have $\mathrm{P}(\tilde{\Omega}) = 1$. This, combined with the fact that

$$\begin{aligned}\sup_{t\in T} \inf_{s\in T_\infty} |X(t)-X(s)| \quad &\leq \quad \sup_{t\in T} \inf_{s\in T_*} |X(t)-X(s)| \\ &\quad + \sup_{t\in T_*} \inf_{s\in T_\infty} |X(s)-X(t)|,\end{aligned}$$

implies that $\sup_{t\in T} \inf_{s\in T_\infty} |X(t)-X(s)| = 0$ almost surely. Thus taking the supremum over $T$ is equivalent to taking the supremum over $T_\infty$.

The second conclusion of the theorem follows from the previous result by setting $\delta = \eta = \mathrm{diam}\, T$ and noting that, in this case, $D(\eta,d) = 1$. Now we have

$$\begin{aligned}\delta\psi^{-1}(D^2(\eta,d)) \quad &= \quad \eta\psi^{-1}(D(\eta,d)) \\ &= \quad \int_0^\eta \psi^{-1}(D(\eta,d))d\epsilon \\ &\leq \quad \int_0^\eta \psi^{-1}(D(\epsilon,d))d\epsilon,\end{aligned}$$

and the second conclusion follows.□

As a consequence of Exercise 8.5.5 below, the conclusions of Theorem 8.4 show that $X$ has $d$-continuous sample paths almost surely whenever the integral $\int_0^\eta \psi^{-1}(D(\epsilon,d))d\epsilon$ is bounded for some $\eta > 0$. It is also easy to verify that the maximum of the process of $X$ is bounded, since $\| \sup_{t\in T} X(t)\|_\psi \leq \|X(t_0)\|_\psi + \| \sup_{s,t\in T} |X(t)-X(s)| \|_\psi$, for any choice of $t_0 \in T$. Thus $X$

is tight and takes its values in $UC(T,d)$ almost surely. These results will prove quite useful in later developments.

An important application of Theorem 8.4 is to *sub-Gaussian* processes:

COROLLARY 8.5 *Let $\{X(t), t \in T\}$ be a separable sub-Gaussian process with respect to $d$. Then for all $\delta > 0$,*

$$\mathrm{E}\left(\sup_{s,t\in T: d(s,t)\le\delta} |X(s) - X(t)|\right) \le K \int_0^\delta \sqrt{\log D(\epsilon, d)}\, d\epsilon,$$

*where $K$ is a universal constant. Also, for any $t_0 \in T$,*

$$\mathrm{E}\left(\sup_{t\in T} |X(t)|\right) \le \mathrm{E}|X(t_0)| + K \int_0^{\operatorname{diam} T} \sqrt{\log D(\epsilon, d)}\, d\epsilon.$$

**Proof.** Apply Theorem 8.4 with $\psi = \psi_2$ and $\eta = \delta$. Because $\psi_2^{-1}(m) = \sqrt{\log(1+m)}$, $\psi_2^{-1}(D^2(\delta, d)) \le \sqrt{2}\psi_2^{-1}(D(\delta, d))$. Hence the second term of the general maximal inequality can be replaced by

$$\sqrt{2}\delta\psi^{-1}(D(\delta, d)) \le \sqrt{2} \int_0^\delta \psi^{-1}(D(\epsilon, d))\, d\epsilon,$$

and we obtain

$$\left\| \sup_{d(s,t)\le\delta} |X(s) - X(t)| \right\|_{\psi_2} \le K \int_0^\delta \sqrt{\log(1 + D(\epsilon, d))}\, d\epsilon,$$

for an enlarged universal constant $K$. Note that $D(\epsilon, d) \ge 2$ for all $\epsilon$ strictly less than $\operatorname{diam} T$. Since $(1 + m) \le m^2$ for all $m \ge 2$, the 1 inside of the logarithm can be removed at the cost of increasing $K$ again, whenever $\delta < \operatorname{diam} T$. Thus it is also true for all $\delta \le \operatorname{diam} T$. We are done with the first conclusion since $d(s,t) \le \operatorname{diam} T$ for all $s, t \in T$. Since the second conclusion is an easy consequence of the first, the proof is complete.□

The next corollary shows how to use the previous corollary to establish bounds on the *modulus of continuity* of certain sub-Gaussian processes. Here the modulus of continuity for a stochastic process $\{X(t) : t \in T\}$, where $(T, d)$ is a semimetric space, is defined as

$$m_X(\delta) \equiv \sup_{s,t\in T: d(s,t)\le\delta} |X(s) - X(t)|.$$

COROLLARY 8.6 *Assume the conditions of Corollary 8.5. Also assume there exists a differentiable function $\delta \mapsto h(\delta)$, with derivative $\dot{h}(\delta)$, satisfying $h(\delta) \ge \sqrt{\log D(\delta, d)}$ for all $\delta > 0$ small enough and $\lim_{\delta\downarrow 0}[\delta\dot{h}(\delta)/h(\delta)] = 0$. Then*

$$\lim_{M\to\infty} \limsup_{\delta\downarrow 0} \mathrm{P}\left(\frac{m_X(\delta)}{\delta h(\delta)} > M\right) = 0.$$

**Proof.** Using L'Hospital's rule and the assumptions of the theorem, we obtain that

$$\frac{\int_0^\delta \sqrt{\log D(\epsilon, d)}d\epsilon}{\delta h(\delta)} \leq \frac{\int_0^\delta h(\epsilon)d\epsilon}{\delta h(\delta)} \to 1,$$

as $\delta \downarrow 0$. The result now follows from the first assertion of Corollary 8.5.□

In the situation where $D(\epsilon, d) \leq K(1/\epsilon)^r$, for constants $0 < r, K < \infty$ and all $\epsilon > 0$ small enough, the above corollary works for $h(\delta) = c\sqrt{\log(1/\delta)}$, for some constant $0 < c < \infty$. This follows from simple calculations. This situation applies, for example, when $X$ is either a standard Brownian motion or a Brownian bridge on $T = [0, 1]$. Both of these processes are sub-Gaussian with respect to the metric $d(s, t) = |s - t|^{1/2}$, and if we let $\eta = \delta^2$, we obtain from the corollary that

$$\lim_{M\to\infty} \limsup_{\eta\downarrow 0} \mathrm{P}\left(\frac{m_X(\eta)}{\sqrt{\eta \log(1/\eta)}} > M\right) = 0.$$

The rate in the denominator is quite precise in this instance since the Lévy modulus theorem (see Theorem 9.25 of Karatzas and Shreve, 1991) yields

$$\mathrm{P}\left(\limsup_{\eta\downarrow 0} \frac{m_X(\eta)}{\sqrt{\eta \log(1/\eta)}} = \sqrt{2}\right) = 1.$$

The above discussion is also applicable to the modulus of continuity of certain empirical processes, and we will examine this briefly in Chapter 11. We note that the condition $\delta\dot{h}(\delta)/h(\delta) \to 0$, as $\delta \downarrow 0$, can be thought of as a requirement that $D(\delta, d)$ goes to infinity extremely slowly as $\delta \downarrow 0$. While this condition is fairly severe, it is reasonably plausible as illustrated in the above Brownian bridge example.

We now consider an important example of a sub-Gaussian process useful for studying empirical processes. This is the *Rademacher process*

$$X(a) = \sum_{i=1}^n \epsilon_i a_i, \quad a \in \mathbb{R}^n,$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. *Rademacher random variables* satisfying $P(\epsilon = -1) = P(\epsilon = 1) = 1/2$. We will verify shortly that this is indeed a sub-Gaussian process with respect to the Euclidean distance $d(a, b) = \|a - b\|$ (which obviously makes $T = \mathbb{R}^n$ into a metric space). This process will emerge in our development of Donsker results based on uniform entropy. The following lemma, also known as Hoeffding's inequality, verifies that Rademacher processes are sub-Gaussian:

LEMMA 8.7 *(Hoeffding's inequality) Let $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$ and $\epsilon_1, \ldots, \epsilon_n$ be independent Rademacher random variables. Then*

$$\mathrm{P}\left(\left|\sum_{i=1}^{n}\epsilon_i a_i\right| > x\right) \le 2e^{-\frac{1}{2}x^2/\|a\|^2},$$

*for the Euclidean norm* $\|\cdot\|$. *Hence* $\|\sum \epsilon a\|_{\psi_2} \le \sqrt{6}\|a\|$.

**Proof.** For any $\lambda$ and Rademacher variable $\epsilon$, one has $\mathrm{E}e^{\lambda\epsilon} = (e^{\lambda} + e^{-\lambda})/2 = \sum_{i=0}^{\infty}\lambda^{2i}/(2i)! \le e^{\lambda^2/2}$, where the last inequality follows from the relation $(2i)! \ge 2^i i!$ for all nonnegative integers. Hence Markov's inequality gives for any $\lambda > 0$

$$\mathrm{P}\left(\sum_{i=1}^{n}\epsilon_i a_i > x\right) \le e^{-\lambda x}\mathrm{E}\exp\left\{\lambda\sum_{i=1}^{n}\epsilon_i a_i\right\} \le \exp\{(\lambda^2/2)\|a\|^2 - \lambda x\}.$$

Setting $\lambda = x/\|a\|^2$ yields the desired upper bound. Since multiplying $\epsilon_1, \ldots, \epsilon_n$ by $-1$ does not change the joint distribution, we obtain

$$\mathrm{P}\left(-\sum_{i=1}^{n}\epsilon_i a_i > x\right) = \mathrm{P}\left(\sum_{i=1}^{n}\epsilon_i a_i > x\right),$$

and the desired upper bound for the absolute value of the sum follows. The bound on the $\psi_2$-norm follows directly from Lemma 8.1.□

## 8.2 The Symmetrization Inequality and Measurability

We now discuss a powerful technique for empirical processes called *symmetrization*. We begin by defining the "symmetrized" empirical process $f \mapsto \mathbb{P}_n^\circ f \equiv n^{-1}\sum_{i=1}^{n}\epsilon_i f(X_i)$, where $\epsilon_1, \ldots, \epsilon_n$ are independent Rademacher random variables which are also independent of $X_1, \ldots, X_n$. The basic idea behind symmetrization is to replace supremums of the form $\|(\mathbb{P}_n - P)f\|_{\mathcal{F}}$ with supremums of the form $\|\mathbb{P}_n^\circ f\|_{\mathcal{F}}$. This replacement is very useful in Glivenko-Cantelli and Donsker theorems based on uniform entropy, and a proof of the validity of this replacement is the primary goal of this section. Note that the processes $(\mathbb{P}_n - P)f$ and $\mathbb{P}_n^\circ f$ both have mean zero. A deeper connection between these two processes is that a Donsker theorem or Glivenko-Cantelli theorem holds for one of these processes if and only if it holds for the other.

One potentially troublesome difficulty is that the supremums involved may not be measurable, and we need to be clear about the underlying product probability spaces so that the outer expectations are well defined. In this setting, we will assume that $X_1, \ldots, X_n$ are the coordinate projections of the product space $(\mathcal{X}^n, \mathcal{A}^n, P^n)$, where $(\mathcal{X}, \mathcal{A}, P)$ is the probability space for a single observation and $\mathcal{A}^n$ is shorthand for

the product $\sigma$-field generated from sets of the form $A_1 \times \cdots \times A_n$, where $A_1, \ldots, A_n \in \mathcal{A}$. In many of the settings of interest to us, the $\sigma$-field $\mathcal{A}^n$ will be strictly smaller than the Borel $\sigma$-field generated from the product topology, as discussed in Section 6.1, but the results we obtain using $\mathcal{A}^n$ will be sufficient for our purposes. In some settings, an additional source of randomness, independent of $X_1, \ldots, X_n$, will be involved which we will denote $Z$. If we let the probability space for $Z$ be $(\mathcal{Z}, \mathcal{D}, Q)$, we will assume that the resulting underlying joint probability space has the form $(\mathcal{X}^n, \mathcal{A}^n, P^n) \times (\mathcal{Z}, \mathcal{D}, Q) = (\mathcal{X}^n \times \mathcal{Z}, \mathcal{A}^n \times \mathcal{D}, P^n \times Q)$, where we define the product $\sigma$-field $\mathcal{A}^n \times \mathcal{D}$ in the same manner as before. Now $X_1, \ldots, X_n$ are equal to the coordinate projections on the first $n$ coordinates, while $Z$ is equal to the coordinate projection on the $(n+1)$st coordinate.

We now present the symmetrization theorem. After its proof, we will discuss a few additional important measurability issues.

THEOREM 8.8 *(Symmetrization) For every nondecreasing, convex $\phi : \mathbb{R} \mapsto \mathbb{R}$ and class of measurable functions $\mathcal{F}$,*

$$\mathrm{E}^* \phi\left(\frac{1}{2}\|\mathbb{P}_n - P\|_{\mathcal{F}}\right) \leq \mathrm{E}^* \phi\left(\|\mathbb{P}_n^\circ\|_{\mathcal{F}}\right) \leq \mathrm{E}^* \phi\left(2\|\mathbb{P}_n - P\|_{\mathcal{F}} + |R_n| \cdot \|P\|_{\mathcal{F}}\right),$$

*where $R_n \equiv \mathbb{P}_n^\circ 1 = n^{-1} \sum_{i=1}^n \epsilon_i$ and the outer expectations are computed based on the product $\sigma$-field described in the previous paragraph.*

Before giving the proof of this theorem, we make a few observations. Firstly, the constants $1/2$, 1 and 2 appearing in front of the three respective supremum norms in the chain of inequalities can all be replaced by $c/2$, $c$ and $2c$, respectively, for any positive constant $c$. This follows trivially since, for any positive $c$, $x \mapsto \phi(cx)$ is nondecreasing and convex whenever $x \mapsto \phi(x)$ is nondecreasing and convex. Secondly, we note that most of our applications of this theorem will be for the setting $\phi(x) = x$. Thirdly, we note that the first inequality in the chain of inequalities will be of greatest use to us. However, the second inequality in the chain can be used to establish the following Glivenko-Cantelli result, the complete proof of which will be given later on, at the tail end of Section 8.3:

PROPOSITION 8.9 *For any class of measurable functions $\mathcal{F}$, the following are equivalent:*

*(i) $\mathcal{F}$ is $P$-Glivenko-Cantelli and $\|P\|_{\mathcal{F}} < \infty$.*

*(ii) $\|\mathbb{P}_n^\circ\|_{\mathcal{F}} \overset{\text{as*}}{\to} 0$.*

As mentioned previously, there is also a similar equivalence involving Donsker results, but we will postpone further discussion of this until we encounter multiplier central limit theorems in Chapter 10.

**Proof of Theorem 8.8.** Let $Y_1, \ldots, Y_n$ be independent copies of $X_1, \ldots, X_n$. Formally, $Y_1, \ldots, Y_n$ are the coordinate projections on the last $n$ coordinates in the product space $(\mathcal{X}^n, \mathcal{A}^n, P^n) \times (\mathcal{Z}, \mathcal{D}, Q) \times (\mathcal{X}^n, \mathcal{A}^n, P^n)$.

Here, $(\mathcal{Z}, \mathcal{D}, Q)$ is the probability space for the $n$-vector of independent Rademacher random variables $\epsilon_1, \ldots, \epsilon_n$ used in $\mathbb{P}_n^\circ$. Since, by Lemma 6.13, coordinate projections are perfect maps, the outer expectations in the theorem are unaffected by the enlarged product probability space. For fixed $X_1, \ldots, X_n$, $\|\mathbb{P}_n - P\|_{\mathcal{F}} =$

$$\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n [f(X_i) - \mathrm{E} f(Y_i)] \right| \leq \mathrm{E}_Y^* \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n [f(X_i) - f(Y_i)] \right|,$$

where $\mathrm{E}_Y^*$ is the outer expectation with respect to $Y_1, \ldots, Y_n$ computed by treating the $X_1, \ldots, X_n$ as constants and using the probability space $(\mathcal{X}^n, \mathcal{A}^n, P^n)$. Applying Jensen's inequality, we obtain

$$\phi\left(\|\mathbb{P}_n - P\|_{\mathcal{F}}\right) \leq \mathrm{E}_Y \phi \left( \left\| \frac{1}{n} \sum_{i=1}^n [f(X_i) - f(Y_i)] \right\|_{\mathcal{F}}^{*Y} \right),$$

where $*Y$ denotes the minimal measurable majorant of the supremum with respect to $Y_1, \ldots, Y_n$ and holding $X_1, \ldots, X_n$ fixed. Because $\phi$ is nondecreasing and continuous, the $*Y$ inside of the $\phi$ in the forgoing expression can be removed after replacing $\mathrm{E}_Y$ with $\mathrm{E}_Y^*$, as a consequence of Lemma 6.8. Now take the expectation of both sides with respect to $X_1, \ldots, X_n$ to obtain

$$\mathrm{E}^* \phi\left(\|\mathbb{P}_n - P\|_{\mathcal{F}}\right) \leq \mathrm{E}_X^* \mathrm{E}_Y^* \phi \left( \frac{1}{n} \left\| \sum_{i=1}^n [f(X_i) - f(Y_i)] \right\|_{\mathcal{F}} \right).$$

The repeated outer expectation can now be bounded above by the joint outer expectation $\mathrm{E}^*$ by Lemma 6.14 (Fubini's theorem for outer expectations).

By the product space structure of the underlying probability space, the outer expectation of any function $g(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ remains unchanged under permutations of its $2n$ arguments. Since $-[f(X_i) - f(Y_i)] = [f(Y_i) - f(X_i)]$, we have for any $n$-vector $(e_1, \ldots, e_n) \in \{-1, 1\}^n$, that $\left\| n^{-1} \sum_{i=1}^n e_i [f(X_i) - f(Y_i)] \right\|_{\mathcal{F}}$ is just a permutation of

$$h(X_1, \ldots, X_n, Y_1, \ldots, Y_n) \equiv \left\| n^{-1} \sum_{i=1}^n [f(X_i) - f(Y_i)] \right\|_{\mathcal{F}}.$$

Hence

$$\mathrm{E}^* \phi\left(\|\mathbb{P}_n - P\|_{\mathcal{F}}\right) \leq \mathrm{E}_\epsilon \mathrm{E}_{X,Y}^* \phi \left\| \frac{1}{n} \sum_{i=1}^n e_i [f(X_i) - f(Y_i)] \right\|_{\mathcal{F}}.$$

Now the triangle inequality combined with the convexity of $\phi$ yields

$$\mathrm{E}^* \phi\left(\|\mathbb{P}_n - P\|_{\mathcal{F}}\right) \leq \mathrm{E}_\epsilon \mathrm{E}_{X,Y}^* \phi\left(2 \|\mathbb{P}_n^\circ\|_{\mathcal{F}}\right).$$

By the perfectness of coordinate projections, $\mathrm{E}^*_{X,Y}$ can be replaced by $\mathrm{E}^*_X\mathrm{E}^*_Y$. Now $\mathrm{E}_\epsilon\mathrm{E}^*_X\mathrm{E}^*_Y$ is bounded above by the joint expectation $\mathrm{E}^*$ by reapplication of Lemma 6.14. This proves the first inequality.

For the second inequality, let $Y_1, \ldots, Y_n$ be independent copies of $X_1, \ldots, X_n$ as before. Holding $X_1, \ldots, X_n$ and $\epsilon_1, \ldots, \epsilon_n$ fixed, we have $\|\mathbb{P}^\circ_n f\|_\mathcal{F} = \|\mathbb{P}^\circ_n(f - Pf) + \mathbb{P}^\circ_n Pf\|_\mathcal{F} =$

$$\|\mathbb{P}^\circ_n(f - \mathrm{E}f(Y)) + R_n Pf\|_\mathcal{F} \le \mathrm{E}^*_Y \left\| \frac{1}{n}\sum_{i=1}^n \epsilon_i [f(X_i) - f(Y_i)] \right\|_\mathcal{F} + |R_n| \cdot \|P\|_\mathcal{F}.$$

Applying Jensen's inequality, we now have

$$\phi\left(\|\mathbb{P}^\circ_n\|_\mathcal{F}\right) \le \mathrm{E}^*_Y \phi\left( \left\| \frac{1}{n}\sum_{i=1}^n \epsilon_i [f(X_i) - f(Y_i)] \right\|_\mathcal{F} + |R_n| \cdot \|P\|_\mathcal{F} \right).$$

Using the permutation argument we used for the first part of the proof, we can replace the $\epsilon_1, \ldots, \epsilon_n$ in the summation with all 1's, and take expectations with respect to $X_1, \ldots, X_n$ and $\epsilon_1, \ldots, \epsilon_n$ (which are still present in $R_n$). This gives us

$$\mathrm{E}^*\phi\left(\|\mathbb{P}^\circ_n\|_\mathcal{F}\right) \le \mathrm{E}_\epsilon\mathrm{E}^*_X\mathrm{E}^*_Y \phi\left( \left\| \frac{1}{n}\sum_{i=1}^n [f(X_i) - f(Y_i)] \right\|_\mathcal{F} + |R_n| \cdot \|P\|_\mathcal{F} \right).$$

After adding and subtracting $Pf$ in the summation and applying the convexity of $\phi$, we can bound the right-hand-side by

$$\begin{aligned} &\frac{1}{2}\mathrm{E}_\epsilon\mathrm{E}^*_X\mathrm{E}^*_Y \phi\left( 2\left\| \frac{1}{n}\sum_{i=1}^n [f(X_i) - Pf] \right\|_\mathcal{F} + |R_n| \cdot \|P\|_\mathcal{F} \right) \\ &\qquad + \frac{1}{2}\mathrm{E}_\epsilon\mathrm{E}^*_X\mathrm{E}^*_Y \phi\left( 2\left\| \frac{1}{n}\sum_{i=1}^n [f(Y_i) - Pf] \right\|_\mathcal{F} + |R_n| \cdot \|P\|_\mathcal{F} \right). \end{aligned}$$

By reapplication of the permutation argument and Lemma 6.14, we obtain the desired upper bound.□

The above symmetrization results will be most useful when the supremum $\|\mathbb{P}^\circ_n\|_\mathcal{F}$ is measurable and Fubini's theorem permits taking the expectation first with respect to $\epsilon_1, \ldots, \epsilon_n$ given $X_1, \ldots, X_n$ and secondly with respect to $X_1, \ldots, X_n$. Without this measurability, only the weaker version of Fubini's theorem for outer expectations applies (Lemma 6.14), and thus the desired reordering of expectations may not be valid. To overcome this difficulty, we will assume that the class $\mathcal{F}$ is a *P-measurable class.* A class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$, on the probability space $(\mathcal{X}, \mathcal{A}, P)$, is $P$-measurable if $(X_1, \ldots, X_n) \mapsto \|\sum_{i=1}^n e_i f(X_i)\|_\mathcal{F}$ is measurable on the completion of $(\mathcal{X}^n, \mathcal{A}^n, P^n)$ for every constant vector $(e_1, \ldots, e_n) \in \mathbb{R}^n$. It is possible to weaken this condition, but at least some measurability

assumptions will usually be needed. In the Donsker theorem for uniform entropy, it will be necessary to assume that several related classes of $\mathcal{F}$ are also $P$-measurable. These additional classes are $\mathcal{F}_\delta \equiv \{f - g : f, g \in \mathcal{F}, \|f - g\|_{P,2} < \delta\}$, for all $\delta > 0$, and $\mathcal{F}_\infty^2 \equiv \{(f - g)^2 : f, g \in \mathcal{F}\}$ (recall that $\|f\|_{P,2} \equiv (Pf^2)^{1/2}$).

Another assumption on $\mathcal{F}$ that is stronger than $P$-measurability and often easier to verify in statistical applications is *pointwise measurability*. A class $\mathcal{F}$ of measurable functions is pointwise measurable if there exists a countable subset $\mathcal{G} \subset \mathcal{F}$ such that for every $f \in \mathcal{F}$, there exists a sequence $\{g_m\} \in \mathcal{G}$ with $g_m(x) \to f(x)$ for every $x \in \mathcal{X}$. Since, by Exercise 8.5.6 below, $\|\sum e_i f(X_i)\|_{\mathcal{F}} = \|\sum e_i f(X_i)\|_{\mathcal{G}}$ for all $(e_1, \ldots, e_n) \in \mathbb{R}^n$, pointwise measurable classes are $P$-measurable for all $P$. Consider, for example, the class $\mathcal{F} = \{1\{x \le t\} : t \in \mathbb{R}\}$ where the sample space $\mathcal{X} = \mathbb{R}$. Let $\mathcal{G} = \{1\{x \le t\} : t \in \mathbb{Q}\}$, and fix the function $x \mapsto f(x) = 1\{x \le t_0\}$ for some $t_0 \in \mathbb{R}$. Note that $\mathcal{G}$ is countable. Let $\{t_m\}$ be a sequence of rationals with $t_m \ge t_0$, for all $m \ge 1$, and with $t_m \downarrow t_0$. Then $x \mapsto g_m(x) = 1\{x \le t_m\}$ satisfies $g_m \in \mathcal{G}$, for all $m \ge 1$, and $g_m(x) \to f(x)$ for all $x \in \mathbb{R}$. Since $t_0$ was arbitrary, we have just proven that $\mathcal{F}$ is pointwise measurable (and hence also $P$-measurable for all $P$). Hereafter, we will use the abbreviation PM as a shorthand for denoting pointwise measurable classes.

Another nice feature of PM classes is that they have a number of useful preservation features. An obvious example is that when $\mathcal{F}_1$ and $\mathcal{F}_2$ are PM classes, then so is $\mathcal{F}_1 \cup \mathcal{F}_2$. The following lemma provides a number of additional preservation results:

LEMMA 8.10 *Let $\mathcal{F}_1, \ldots, \mathcal{F}_k$ be PM classes of real functions on $\mathcal{X}$, and let $\phi : \mathbb{R}^k \mapsto \mathbb{R}$ be continuous. Then the class $\phi \circ (\mathcal{F}_1, \ldots, \mathcal{F}_k)$ is PM, where $\phi \circ (\mathcal{F}_1, \ldots, \mathcal{F}_k)$ denotes the class $\{\phi(f_1, \ldots, f_k) : (f_1, \ldots, f_k) \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_k\}$.*

**Proof.** Denote $\mathcal{H} \equiv \phi \circ (\mathcal{F}_1, \ldots, \mathcal{F}_k)$. Fix an arbitrary $h = \phi(f_1, \ldots, f_k) \in \mathcal{H}$. By assumption, each $\mathcal{F}_j$ has a countable subset $\mathcal{G}_j \subset \mathcal{F}_j$ such that there exists a subsequence $\{g_m^j\} \in \mathcal{G}_j$ with $g_m^j(x) \to f_j(x)$, as $m \to \infty$, for all $x \in \mathcal{X}$ and $j = 1, \ldots, k$. By continuity of $\phi$, we thus have that $\phi(g_m^1(x), \ldots, g_m^k(x)) \to \phi(f_1(x), \ldots, f_k(x)) = h(x)$, as $m \to \infty$, for all $x \in \mathcal{X}$. Since the choice of $h$ was arbitrary, we therefore have that the set $\phi(\mathcal{G}_1, \ldots, \mathcal{G}_k)$ is a countable subset of $\mathcal{H}$ making $\mathcal{H}$ pointwise measurable.□

Lemma 8.10 automatically yields many other useful PM preservation results, including the following for PM classes $\mathcal{F}_1$ and $\mathcal{F}_2$:

- $\mathcal{F}_1 \wedge \mathcal{F}_2$ (all possible pairwise minimums) is PM.
- $\mathcal{F}_1 \vee \mathcal{F}_2$ (all possible pairwise maximums) is PM.
- $\mathcal{F}_1 + \mathcal{F}_2$ is PM.
- $\mathcal{F}_1 \cdot \mathcal{F}_2 \equiv \{f_1 f_2 : f_1 \in \mathcal{F}_1,\ f_2 \in \mathcal{F}_2\}$ is PM.

We will use these properties of PM classes to establish Donsker properties for some specific statistical examples later on in the case studies presented in Chapter 15. The following proposition shows an additional property of PM classes that potentially simplifies the measurability requirements of the Donsker theorem for uniform entropy, Theorem 8.19, given in Section 8.4 below:

PROPOSITION 8.11 *Let $\mathcal{F}$ be a class of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$ on the probability space $(\mathcal{X}, \mathcal{A}, P)$. Provided $\mathcal{F}$ is PM with envelope $F$ such that $P^*F^2 < \infty$, then $\mathcal{F}_\delta$ and $\mathcal{F}_\infty^2$ are PM for all $0 < \delta \leq \infty$.*

**Proof.** The fact that both $\mathcal{F}_\infty$ and $\mathcal{F}_\infty^2$ are PM follows easily from Lemma 8.10. Assume, without loss of generality, that the envelope $F$ is measurable (if not, simply replace $F$ with $F^*$). Next, let $\mathcal{H} \subset \mathcal{F}_\infty$ be a countable subset for which there exists for each $g \in \mathcal{F}_\infty$ a sequence $\{h_m\} \in \mathcal{H}$ such that $h_m(x) \to g(x)$ for all $x \in \mathcal{X}$. Fix $\delta > 0$ and $h \in \mathcal{F}_\delta$. Then there exists an $\epsilon > 0$ such that $Ph^2 = \delta^2 - \epsilon$. Let $\{g_m\} \in \mathcal{H}$ be a sequence for which $g_m(x) \to h(x)$ for all $x \in \mathcal{X}$, and assume that $Pg_m^2 \geq \delta^2$ infinitely often. Then there exists another sequence $\{\tilde{g}_m\} \in \mathcal{H}$ such that $P\tilde{g}_m^2 \geq \delta^2$ for all $m \geq 1$ and also $\tilde{g}_m(x) \to h(x)$ for all $x \in \mathcal{X}$. Since $|\tilde{g}_m| \leq F$, for all $m \geq 1$, we have by the dominated convergence theorem that $\delta^2 \leq \liminf_{m\to\infty} P\tilde{g}_m^2 = Ph^2 = \delta^2 - \epsilon$, which is impossible. Hence, returning to the original sequence $\{g_m\}$, $\|g_m\|_{P,2}$ cannot be $\geq \delta$ infinitely often. Thus there exists a sequence $\{\check{g}_m\} \in \mathcal{H}_\delta \equiv \{g \in \mathcal{H} : \|g\|_{P,2} < \delta\}$ such that $\check{g}_m(x) \to h(x)$ for all $x \in \mathcal{X}$. Thus $\mathcal{F}_\delta$ is PM since $h$ was arbitrary and $\mathcal{H}_\delta$ does not depend on $h$. Since $\delta$ was also arbitrary, the proof is complete.□

We next consider establishing $P$-measurability for the class

$$\left\{1\{Y - \beta^T Z \leq t\} : \beta \in \mathbb{R}^k, t \in \mathbb{R}\right\},$$

where $X \equiv (Y, Z) \in \mathcal{X} \equiv \mathbb{R} \times \mathbb{R}^k$ has distribution $P$, for arbitrary $P$. This class was considered in the linear regression example of Section 4.1. The desired measurability result is stated in the following lemma:

LEMMA 8.12 *Let $\mathcal{F} \equiv \left\{1\{Y - \beta^T Z \leq t\} : \beta \in \mathbb{R}^k, t \in \mathbb{R}\right\}$. Then the classes $\mathcal{F}$, $\mathcal{F}_\delta \equiv \{f - g : f, g \in \mathcal{F}, \|f - g\|_{P,2} < \delta\}$, and $\mathcal{F}_\infty^2 \equiv \left\{(f - g)^2 : f, g \in \mathcal{F}\right\}$ are all $P$-measurable for any probability measure on $\mathcal{X}$.*

**Proof.** We first assume that $\|Z\| \leq M$ for some fixed $M < \infty$. Hence the sample space is $\mathcal{X}_M \equiv \{(y, z) : (y, z) \in \mathcal{X}, \|z\| \leq M\}$. Consider the countable set $\mathcal{G} = \left\{1\{Y - \beta^T Z \leq t\} : \beta \in \mathbb{Q}^k, t \in \mathbb{Q}\right\}$, where $\mathbb{Q}$ are the rationals. Fix $\beta \in \mathbb{R}^k$ and $t \in \mathbb{R}$, and construct a sequence $\{(\beta_m, t_m)\}$ as follows: for each $m \geq 1$, pick $\beta_m \in \mathbb{Q}^k$ so that $\|\beta_m - \beta\| < 1/(2mM)$ and pick $t_m \in \mathbb{Q}$ so that $t_m \in (t + 1/(2m), t + 1/m]$. Now, for any $(y, z)$ with $y \in \mathbb{R}$ and $z \in \mathbb{R}^k$ with $\|z\| \leq M$, we have that $1\{y - \beta_m^T z \leq t_m\} = 1\{y - \beta^T z \leq t_m + (\beta_m - \beta)^T z\}$. Since $|(\beta_m - \beta)^T z| < 1/(2m)$ by design,

we have that $r_m \equiv t_m + (\beta_m - \beta)^T z - t > 0$ for all $m$ and that $r_m \to 0$ as $m \to \infty$. Since the function $t \mapsto 1\{u \le t\}$ is right-continuous and since $(y, z)$ was arbitrary, we have just proven that $1\{y - \beta_m^T z \le t_m\} \to \{y - \beta^T z \le t\}$ for all $(y, z) \in \mathcal{X}_M$. Thus $\mathcal{F}$ is pointwise measurable with respect to the countable subset $\mathcal{G}$.

We can also verify that $\mathcal{F}_\delta$ and $\mathcal{F}_\infty^2$ are likewise PM classes, under the constraint that the random variable $Z$ satisfies $\|Z\| \le M$. To see this for $\mathcal{F}_\delta$, let $f_1, f_2 \in \mathcal{F}$ satisfy $\|f_1 - f_2\|_{P,2} < \delta$ and let $\{g_{m,1}\}, \{g_{m,2}\} \in \mathcal{G}$ be such that $g_{m,1} \to f_1$ and $g_{m,2} \to f_2$ pointwise in $\mathcal{X}_M$. Then, by dominated convergence, $\|g_{m,1} - g_{m,2}\|_{P,2} \to \|f_1 - f_2\|_{P,2}$, and thus $\|g_{m,1} - g_{m,2}\|_{P,2} < \delta$ for all $m$ large enough. Hence

$$g_{m,1} - g_{m,2} \in \mathcal{G}_\delta \equiv \{f - g : f, g \in \mathcal{G},\ \|f - g\|_{P,2} < \delta\}$$

for all $m$ large enough, and thus $\mathcal{G}_\delta$ is a separable subset of $\mathcal{F}_\delta$ making $\mathcal{F}_\delta$ into a PM class. The proof that $\mathcal{F}_\infty^2$ is also PM follows directly from Lemma 8.10.

Now let $J_M(x_1, \ldots, x_n) = 1\{\max_{1 \le i \le n} \|z_i\| \le M\}$, where $x_i = (y_i, z_i)$, $1 \le i \le n$. Since $M$ was arbitrary, the previous two paragraphs have established that

$$(x_1, \ldots, x_n) \quad \mapsto \quad \left\| \sum_{i=1}^n e_i f(x_i) \right\|_{\mathcal{H}} J_M(x_1, \ldots, x_n) \tag{8.7}$$

is measurable for every $n$-tuple $(e_1, \ldots, e_n) \in \mathbb{R}^n$, every $M < \infty$, and with $\mathcal{H}$ being replaced by $\mathcal{F}$, $\mathcal{F}_\delta$ or $\mathcal{F}_\infty^2$. Now, for any $(x_1, \ldots, x_n) \in \mathcal{X}^n$, $J_M(x_1, \ldots, x_n) = 1$ for all $M$ large enough. Thus the map (8.7) is also measurable after replacing $J_M$ with its pointwise limit $1 = \lim_{M \to \infty} J_M$. Hence $\mathcal{F}$, $\mathcal{F}_\delta$ and $\mathcal{F}_\infty^2$ are all $P$-measurable classes for any measure $P$ on $\mathcal{X}$.□

Another example of a $P$-measurable class occurs when $\mathcal{F}$ is a Suslin topological space (for an arbitrary topology $\mathcal{O}$), 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 generated from $\mathcal{O}$. Further insights and results on this *measurable Suslin condition* can be found in Example 2.3.5 and Chapter 1.7 of VW. While this approach to establishing measurability can be useful in some settings, a genuine need for it does not often occur in statistical applications, and we will not pursue it further here.

## 8.3 Glivenko-Cantelli Results

We now present several Glivenko-Cantelli (G-C) results. First, we discuss an interesting necessary condition for a class $\mathcal{F}$ to be $P$-G-C. Next, we present the proofs of G-C theorems for bracketing (Theorem 2.2 on Page 16)

and uniform (Theorem 2.4 on Page 18) entropy. Part of the proof in the uniform entropy case will include the presentation of a new G-C theorem, Theorem 8.15 below. Finally, we give the proof of Proposition 8.9 which was promised in the previous section.

The following lemma shows that the existence of an integrable envelope of the centered functions of a class $\mathcal{F}$ is a necessary condition for $\mathcal{F}$ to be $P$-G-C:

LEMMA 8.13 *If the class of functions $\mathcal{F}$ is strong $P$-G-C, then $P\|f - Pf\|^*_{\mathcal{F}} < \infty$. If in addition $\|P\|_{\mathcal{F}} < \infty$, then also $P\|f\|^*_{\mathcal{F}} < \infty$.*

**Proof.** Since $f(X_n) - Pf = n(\mathbb{P}_n - P)f - (n-1)(\mathbb{P}_{n-1} - P)f$, we have $n^{-1}\|f(X_n) - Pf\|_{\mathcal{F}} \leq \|\mathbb{P}_n - P\|_{\mathcal{F}} + (1 - n^{-1})\|\mathbb{P}_{n-1} - P\|_{\mathcal{F}}$. Since $\mathcal{F}$ is strong $P$-G-C, we now have that $P(\|f(X_n) - Pf\|^*_{\mathcal{F}} \geq n \text{ infinitely often}) = 0$. The Borel-Cantelli lemma now yields that $\sum_{n=1}^{\infty} P(\|f(X_n) - Pf\|^*_{\mathcal{F}} \geq n) < \infty$. Since the $X_n$ are i.i.d., the $f(X_n)$ in the summands can be replaced with $f(X_1)$ for all $n \geq 1$. Now we have

$$\begin{aligned} P^*\|f - Pf\|_{\mathcal{F}} &\leq \int_0^{\infty} P\left(\|f(X_1) - Pf\|^*_{\mathcal{F}} > x\right) dx \\ &\leq 1 + \sum_{n=1}^{\infty} P\left(\|f(X_1) - Pf\|^*_{\mathcal{F}} \geq n\right) \\ &< \infty. \square \end{aligned}$$

**Proof of Theorem 2.2 (see Page 16).** Fix $\epsilon > 0$. Since the $L_1$-bracketing entropy is bounded, it is possible to choose finitely many $\epsilon$-brackets $[l_i, u_i]$ so that their union contains $\mathcal{F}$ and $P(u_i - l_i) < \epsilon$ for every $i$. Now, for every $f \in \mathcal{F}$, there is a bracket $[l_i, u_i]$ containing $f$ with $(\mathbb{P}_n - P)f \leq (\mathbb{P}_n - P)u_i + P(u_i - f) \leq (\mathbb{P}_n - P)u_i + \epsilon$. Hence

$$\begin{aligned} \sup_{f \in \mathcal{F}} (\mathbb{P}_n - P)f &\leq \max_i (\mathbb{P}_n - P)u_i + \epsilon \\ &\overset{\text{as}*}{\to} \epsilon. \end{aligned}$$

Similar arguments can be used to verify that

$$\begin{aligned} \inf_{f \in \mathcal{F}} (\mathbb{P}_n - P)f &\geq \min_i (\mathbb{P}_n - P)l_i - \epsilon \\ &\overset{\text{as}*}{\to} -\epsilon. \end{aligned}$$

The desired result now follows since $\epsilon$ was arbitrary.$\square$

To prove Theorem 2.4 on Page 18, we first restate the theorem to clarify the meaning of "appropriately measurable" in the original statement of the theorem, and then prove a more general version (Theorem 8.15 below):

THEOREM 8.14 *(Restated Theorem 2.4 from Page 18) Let $\mathcal{F}$ be a $P$-measurable class of measurable functions with envelope $F$ and*

$$\sup_Q N(\epsilon \|F\|_{Q,1}, \mathcal{F}, L_1(Q)) < \infty,$$

*for every $\epsilon > 0$, where the supremum is taken over all finite probability measures $Q$ with $\|F\|_{Q,1} > 0$. If $P^*F < \infty$, then $\mathcal{F}$ is P-G-C.*

**Proof.** The result is trivial if $P^*F = 0$. Hence we will assume without loss of generality that $P^*F > 0$. Thus there exists an $\eta > 0$ such that, with probability 1, $\mathbb{P}_n F > \eta$ for all $n$ large enough. Fix $\epsilon > 0$. By assumption, there is a $K < \infty$ such that $1\{\mathbb{P}_n F > 0\} \log N(\epsilon \mathbb{P}_n F, \mathcal{F}, L_1(\mathbb{P}_n)) \le K$ almost surely, since $\mathbb{P}_n$ is a finite probability measure. Hence, with probability 1, $\log N(\epsilon\eta, \mathcal{F}, L_1(\mathbb{P}_n)) \le K$ for all $n$ large enough. Since $\epsilon$ was arbitrary, we now have that $\log N(\epsilon, \mathcal{F}, L_1(\mathbb{P}_n)) = O_P^*(1)$ for all $\epsilon > 0$. Now fix $\epsilon > 0$ (again) and $M < \infty$, and define $\mathcal{F}_M \equiv \{f 1\{F \le M\} : f \in \mathcal{F}\}$. Since, $\|(f-g)1\{F \le M\}\|_{1,\mathbb{P}_n} \le \|f-g\|_{1,\mathbb{P}_n}$ for any $f, g \in \mathcal{F}$, $N(\epsilon, \mathcal{F}_M, L_1(\mathbb{P}_n)) \le N(\epsilon, \mathcal{F}, L_1(\mathbb{P}_n))$. Hence $\log N(\epsilon, \mathcal{F}_M, L_1(\mathbb{P}_n)) = O_P^*(1)$. Finally, since $\epsilon$ and $M$ are both arbitrary, the desired result follows from Theorem 8.15 below.□

THEOREM 8.15 *Let $\mathcal{F}$ be a P-measurable class of measurable functions with envelope $F$ such that $P^*F < \infty$. Let $\mathcal{F}_M$ be as defined in the above proof. If $\log N(\epsilon, \mathcal{F}_M, L_1(\mathbb{P}_n)) = o_P^*(n)$ for every $\epsilon > 0$ and $M < \infty$, then $P\|\mathbb{P}_n - P\|_{\mathcal{F}}^* \to 0$ and $\mathcal{F}$ is strong P-G-C.*

Before giving the proof of Theorem 8.15, we give the following lemma which will be needed. This is Lemma 2.4.5 of VW, and we omit the proof:

LEMMA 8.16 *Let $\mathcal{F}$ be a class of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$ with integrable envelope. 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 $\mathrm{E}(\|\mathbb{P}_n - P\|_{\mathcal{F}}^* | \Sigma_{n+1}) \ge \|\mathbb{P}_{n+1} - P\|_{\mathcal{F}}^*$, almost surely. 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 of Theorem 8.15.** By the symmetrization Theorem 8.8, $P$-measura- bility of $\mathcal{F}$, and by Fubini's theorem, we have for all $M > 0$ that

$$\begin{aligned}
\mathrm{E}^*\|\mathbb{P}_n - P\|_{\mathcal{F}} &\le 2\mathrm{E}_X\mathrm{E}_\epsilon \left\| \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) \right\|_{\mathcal{F}} \\
&\le 2\mathrm{E}_X\mathrm{E}_\epsilon \left\| \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) 1\{F \le M\} \right\|_{\mathcal{F}} \\
&\quad + 2\mathrm{E}_X\mathrm{E}_\epsilon \left\| \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) 1\{F > M\} \right\|_{\mathcal{F}} \\
&\le 2\mathrm{E}_X\mathrm{E}_\epsilon \left\| \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) \right\|_{\mathcal{F}_M} + 2P^*\left[F 1\{F > M\}\right].
\end{aligned}$$

The last term can be made arbitrarily small by making $M$ large enough. Thus, for convergence in mean, it is enough to show that the first term goes to zero for each $M$. Accordingly, fix $M < \infty$. Fix also $X_1, \ldots, X_n$, and let $\mathcal{G}$ be a finite $\delta$-mesh in $L_1(\mathbb{P}_n)$ over $\mathcal{F}_M$. Thus

$$\mathrm{E}_\epsilon \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \right\|_{\mathcal{F}_M} \leq \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \right\|_{\mathcal{G}} + \delta. \tag{8.8}$$

By definition of the entropy number, the cardinality of $\mathcal{G}$ can be chosen equal to $N(\delta, \mathcal{F}_M, L_1(\mathbb{P}_n))$. Now, we can bound the $L_1$-norm on the right-hand-side of (8.8) by the Orlicz-norm for $\psi_2(x) = \exp(x^2) - 1$, and apply the maximal inequality Lemma 8.2 to find that the left-hand-side of (8.8) is bounded by a universal constant times

$$\sqrt{1 + \log N(\delta, \mathcal{F}_M, L_1(\mathbb{P}_n))} \sup_{f \in \mathcal{F}} \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \right\|_{\psi_2 | X} + \delta,$$

where the Orlicz norms $\| \cdot \|_{\psi_2 | X}$ are taken over $\epsilon_1, \ldots, \epsilon_n$ with $X_1, \ldots, X_n$ still fixed. From Exercise 8.5.7 below, we have—by Hoeffding's inequality (Lemma 8.7) combined with Lemma 8.1—that the Orlicz norms are all bounded by $\sqrt{6/n}\,(\mathbb{P}_n f^2)^{1/2}$, which is bounded by $\sqrt{6/n}\,M$. The last displayed expression is thus bounded by

$$\sqrt{\frac{6\{1 + \log N(\delta, \mathcal{F}_M, L_1(\mathbb{P}_n))\}}{n}} M + \delta \xrightarrow{\mathrm{P}} \delta.$$

Thus the left-hand-side of (8.8) goes to zero in probability. Since it is also bounded by $M$, the bounded convergence theorem implies that its expectation also goes to zero. Since $M$ was arbitrary, we now have that $\mathrm{E}^* \|\mathbb{P}_n - P\|_{\mathcal{F}} \to 0$. This now implies that $\mathcal{F}$ is weak $P$-G-C.

From Lemma 8.16, we know that there exists a version of $\|\mathbb{P}_n - P\|^*_{\mathcal{F}}$ that converges almost surely to a constant. Since we already know that $\|\mathbb{P}_n - P\|^*_{\mathcal{F}} \xrightarrow{\mathrm{P}} 0$, this constant must be zero. The desired result now follows.□

**Proof of Proposition 8.9.** Assume (i). By the second inequality of the symmetrization theorem (Theorem 8.8), $\|\mathbb{P}^\circ_n\|_{\mathcal{F}} \xrightarrow{\mathrm{P}} 0$. This convergence can be strengthened to outer almost surely, since $\|\mathbb{P}^\circ_n\|_{\mathcal{F}}$ forms a reverse submartingale as in the previous proof. Now assume (ii). By Lemma 8.13 and the fact that $P[\epsilon f(X)] = 0$ for a Rademacher $\epsilon$ independent of $X$, we obtain that $P\|f\|^*_{\mathcal{F}} = P\|\epsilon f(X)\|^*_{\mathcal{F}} < \infty$. Now, the fact that $\mathcal{F}$ is weak $P$-G-C follows from the first inequality in the symmetrization theorem. The convergence can be strengthened to outer almost sure by the reverse martingale argument used previously. Thus (ii) follows.□

## 8.4 Donsker Results

We now present several Donsker results. We begin with several interesting necessary and sufficient conditions for a class to be $P$-Donsker. We next present the proofs of Donsker theorems for bracketing (Theorem 2.3 on Page 17) and uniform (Theorem 2.5 on Page 18) entropy. Before proceeding, let $\mathcal{F}_\delta \equiv \{f - g : f, g \in \mathcal{F}, \|f - g\|_{P,2} < \delta\}$, for any $0 < \delta \leq \infty$.

The following lemma outlines several properties of Donsker classes and shows that Donsker classes are automatically strong Glivenko-Cantelli classes:

LEMMA 8.17 *Let $\mathcal{F}$ be a class of measurable functions, with envelope $F \equiv \|f\|_{\mathcal{F}}$. For any $f, g \in \mathcal{F}$, define $\rho(f,g) \equiv \left\{P(f - Pf - g + Pg)^2\right\}^{1/2}$; and, for any $\delta > 0$, let $\mathcal{F}_\delta \equiv \{f - g : \rho(f,g) < \delta\}$. Then the following are equivalent:*

*(i) $\mathcal{F}$ is $P$-Donsker;*

*(ii) $(\mathcal{F}, \rho)$ is totally bounded and $\|\mathbb{G}_n\|_{\mathcal{F}_{\delta_n}} \xrightarrow{\mathrm{P}} 0$ for every $\delta_n \downarrow 0$;*

*(iii) $(\mathcal{F}, \rho)$ is totally bounded and $\mathrm{E}^*\|\mathbb{G}_n\|_{\mathcal{F}_{\delta_n}} \to 0$ for every $\delta_n \downarrow 0$.*

*These conditions imply that $\mathrm{E}^*\|\mathbb{G}_n\|_{\mathcal{F}}^r \to \mathrm{E}\|\mathbb{G}\|_{\mathcal{F}}^r < \infty$, for every $0 < r < 2$; that $P\left(\|f - Pf\|_{\mathcal{F}}^* > x\right) = o(x^{-2})$ as $x \to \infty$; and that $\mathcal{F}$ is strong $P$-G-C. If in addition $\|P\|_{\mathcal{F}} < \infty$, then also $P(F^* > x) = o(x^{-2})$ as $x \to \infty$.*

**Proof.** The equivalence of (i)–(iii) and the first assertion is Lemma 2.3.11 of VW, and we omit the equivalence part of the proof. Now assume Conditions (i)–(iii) hold. Lemma 2.3.9 of VW states that if $\mathcal{F}$ is Donsker, then

$$\lim_{x \to \infty} x^2 \sup_{n \geq 1} \mathrm{P}^*\left(\|\mathbb{G}_n\|_{\mathcal{F}} > x\right) \quad = \quad 0. \tag{8.9}$$

This immediately implies that the $r$th moment of $\|\mathbb{G}_n\|_{\mathcal{F}}$ is uniformly bounded in $n$ and that $\mathrm{E}\|\mathbb{G}\|_{\mathcal{F}}^r < \infty$ for all $0 < r < 2$. Thus the first assertion follows and, therefore, $\mathcal{F}$ is weak $P$-G-C. Lemma 8.16 now implies $\mathcal{F}$ is strong G-C. Letting $n = 1$ in (8.9) yields that $P\left(\|f - Pf\|_{\mathcal{F}}^* > x\right) = o(x^{-2})$ as $x \to \infty$, and the remaining assertions follow.□

**Proof of Theorem 2.3 (Donsker with bracketing entropy, on Page 17).** With a given set of $\epsilon/2$-brackets $[l_i, u_i]$ covering $\mathcal{F}$, we can construct a set of $\epsilon$-brackets covering $\mathcal{F}_\infty$ by taking differences $[l_i - u_j, u_i - l_j]$ of upper and lower bounds, i.e., if $f \in [l_i, u_i]$ and $g \in [l_j, u_j]$, then $f - g \in [l_i - u_j, u_i - l_j]$. Thus $N_{[]}(\epsilon, \mathcal{F}_\infty, L_2(P)) \leq N_{[]}^2(\epsilon/2, \mathcal{F}, L_2(P))$. From Exercise 8.5.8 below, we now have that $J_{[]}(\delta, \mathcal{F}_\infty, L_2(P)) \leq \sqrt{8} J_{[]}(\delta, \mathcal{F}, L_2(P))$.

Now, for a given, small $\delta > 0$, select a minimal number of $\delta$-brackets that cover $\mathcal{F}$, and use them to construct a finite partition $\mathcal{F} = \cup_{i=1}^m \mathcal{F}_i$ consisting of sets of $L_2(P)$-diameter $\delta$. For any $i \in \{1, \ldots, m\}$, the subset

of $\mathcal{F}_\infty$ consisting of all $f-g$ with $f,g \in \mathcal{F}_i$ consists of functions with $L_2(P)$ norm strictly smaller than $\delta$. Hence by Lemma 8.18 below, there exists a number $a(\delta) > 0$ satisfying

$$\text{(8.10)} \quad \mathrm{E}^* \left[ \sup_{1 \le i \le m} \sup_{f,g \in \mathcal{F}_i} |\mathbb{G}_n(f-g)| \right] \lesssim J_{[]}(\delta, \mathcal{F}, L_2(P)) + \sqrt{n} P^* \left[ G 1\{G > a(\delta)\sqrt{n}\} \right],$$

where $G$ is an envelope for $\mathcal{F}_\infty$ and the relation $\lesssim$ means that the left-hand-side is bounded above by a universal positive constant times the right-hand-side. In this setting, "universal" means that the constant does not depend on $n$ or $\delta$. If $[l,u]$ is a minimal bracket for covering all of $\mathcal{F}$, then $G$ can be taken to be $u-l$. Boundedness of the entropy integral implies that there exists some $k < \infty$ so that only one $L_2(P)$ bracket of size $k$ is needed to cover $\mathcal{F}$. This implies $PG^2 < \infty$.

By Exercise 8.5.9 below, the second term on the right-hand-side of (8.10) is bounded above by $[a(\delta)]^{-1} P^* \left[ G^2 1\{G > a(\delta)\sqrt{n}\} \right]$ and thus goes to zero as $n \to \infty$. Since $J_{[]}(\delta, \mathcal{F}, L_2(P)) = o(\delta)$, as $\delta \downarrow 0$, we now have that $\lim_{\delta \downarrow 0} \limsup_{n \to \infty}$ of the left-hand-side of (8.10) goes to zero. In view of Markov's inequality for outer probability (which follows from Chebyshev's inequality for outer probability as given in Lemma 6.10), Condition (ii) in Lemma 7.20 is satisfied for the stochastic process $X_n(f) = \mathbb{G}_n(f)$ with index set $T = \mathcal{F}$. Now, Theorem 2.1 yields the desired result.□

LEMMA 8.18 *For any class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$ with $Pf^2 < \delta^2$ for every $f$, we have, with*

$$a(\delta) \equiv \frac{\delta}{\sqrt{1 \vee \log N_{[]}(\delta, \mathcal{F}, L_2(P))}},$$

*and $F$ an envelope function for $\mathcal{F}$, that*

$$\mathrm{E}^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J_{[]}(\delta, \mathcal{F}, L_2(P)) + \sqrt{n} P^* \left[ F 1\{F > \sqrt{n} a(\delta)\} \right].$$

This is Lemma 19.34 of van der Vaart (1998), who gives a nice proof which utilizes the maximal inequality result given in Lemma 8.3 above. The arguments are lengthy, and we omit the proof.

To prove Theorem 2.5 (Donsker with uniform entropy, from Page 18), we first restate the theorem with a clarification of the measurability assumption, as done in the previous section for Theorem 2.4:

THEOREM 8.19 *(Restated Theorem 2.5 from Page 18) Let $\mathcal{F}$ be a class of measurable functions with envelope $F$ and $J(1, \mathcal{F}, L_2) < \infty$. Let the classes $\mathcal{F}_\delta$ and $\mathcal{F}_\infty^2 \equiv \{h^2 : h \in \mathcal{F}_\infty\}$ be $P$-measurable for every $\delta > 0$. If $P^* F^2 < \infty$, then $\mathcal{F}$ is $P$-Donsker.*

We note here that by Proposition 8.11, if $\mathcal{F}$ is PM, then so are $\mathcal{F}_\delta$ and $\mathcal{F}_\infty^2$, for all $\delta > 0$, provided $\mathcal{F}$ has envelope $F$ such that $P^*F^2 < \infty$. Since PM implies $P$-measurability, all measurability requirements for Theorem 8.19 are thus satisfied whenever $\mathcal{F}$ is PM.

**Proof of Theorem 8.19.** Let the positive, decreasing sequence $\delta_n \downarrow 0$ be arbitrary. By Markov's inequality for outer probability (see Lemma 6.10) and the symmetrization Theorem 8.8,

$$\mathrm{P}^*\left(\|\mathbb{G}_n\|_{\mathcal{F}_{\delta_n}} > x\right) \le \frac{2}{x}\mathrm{E}^*\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(X_i)\right\|_{\mathcal{F}_{\delta_n}},$$

for i.i.d. Rademachers $\epsilon_1, \ldots, \epsilon_n$ independent of $X_1, \ldots, X_n$. By the $P$-measurability assumption for $\mathcal{F}_\delta$, for all $\delta > 0$, the standard version of Fubini's theorem applies, and the outer expectation is just a standard expectation and can be calculated in the order $\mathrm{E}_X\mathrm{E}_\epsilon$. Accordingly, fix $X_1, \ldots, X_n$. By Hoeffding's inequality (Lemma 8.7), the stochastic process $f \mapsto n^{-1/2} \times \sum_{i=1}^n \epsilon_i f(X_i)$ is sub-Gaussian for the $L_2(\mathbb{P}_n)$-seminorm

$$\|f\|_n \equiv \sqrt{\frac{1}{n}\sum_{i=1}^n f^2(X_i)}.$$

This stochastic process is also separable since, for any measure $Q$ and $\epsilon > 0$, $N(\epsilon, \mathcal{F}_{\delta_n}, L_2(Q)) \le N(\epsilon, \mathcal{F}_\infty, L_2(Q)) \le N^2(\epsilon/2, \mathcal{F}, L_2(Q))$, and the latter is finite for any finite dimensional probability measure $Q$ and any $\epsilon > 0$. Thus the second conclusion of Corollary 8.5 holds with

$$(8.11)\quad \mathrm{E}_\epsilon\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(X_i)\right\|_{\mathcal{F}_{\delta_n}} \lesssim \int_0^\infty \sqrt{\log N(\epsilon, \mathcal{F}_{\delta_n}, L_2(\mathbb{P}_n))}\,d\epsilon.$$

Note that we can omit the term $\mathrm{E}\left|n^{-1/2}\sum_{i=1}^n \epsilon_i f_0(X_i)\right|$ from the conclusion of the corollary because it is also bounded by the right-hand-side of (8.11).

For sufficiently large $\epsilon$, the set $\mathcal{F}_{\delta_n}$ fits in a single ball of $L_2(\mathbb{P}_n)$-radius $\epsilon$ around the origin, in which case the integrand on the right-hand-side of (8.11) is zero. This will definitely happen when $\epsilon$ is larger than $\theta_n$, where

$$\theta_n^2 \equiv \sup_{f\in\mathcal{F}_{\delta_n}} \|f\|_n^2 = \left\|\frac{1}{n}\sum_{i=1}^n f^2(X_i)\right\|_{\mathcal{F}_{\delta_n}}.$$

Thus the right-hand-side of (8.11) is bounded by

$$\begin{aligned}\int_0^{\theta_n} \sqrt{\log N(\epsilon, \mathcal{F}_{\delta_n}, L_2(\mathbb{P}_n))}d\epsilon \\ &\lesssim \int_0^{\theta_n} \sqrt{\log N^2(\epsilon/2, \mathcal{F}, L_2(\mathbb{P}_n))}d\epsilon \\ &\lesssim \int_0^{\theta_n/(2\|F\|_n)} \sqrt{\log N(\epsilon\|F\|_n, \mathcal{F}, L_2(\mathbb{P}_n))}d\epsilon\|F\|_n \\ &\lesssim \|F\|_n J(\theta_n, \mathcal{F}, L_2).\end{aligned}$$

The second inequality follows from the change of variables $u = \epsilon/(2\|F\|_n)$ (and then renaming $u$ to $\epsilon$). For the third inequality, note that we can add $1/2$ to the envelope function $F$ without changing the existence of its second moment. Hence $\|F\|_n \geq 1/2$ without loss of generality, and thus $\theta_n/(2\|F\|_n) \leq \theta_n$. Because $\|F\|_n = O_p(1)$, we can now conclude that the left-hand-side of (8.11) goes to zero in probability, provided we can verify that $\theta_n \xrightarrow{\mathrm{P}} 0$. This would then imply asymptotic $L_2(P)$-equicontinuity in probability.

Since $\|Pf^2\|_{\mathcal{F}_{\delta_n}} \to 0$ and $\mathcal{F}_{\delta_n} \subset \mathcal{F}_\infty$, establishing that $\|\mathbb{P}_n - P\|_{\mathcal{F}_\infty^2} \xrightarrow{\mathrm{P}} 0$ would prove that $\theta_n \xrightarrow{\mathrm{P}} 0$. The class $\mathcal{F}_\infty^2$ has integrable envelope $(2F)^2$ and is $P$-measurable by assumption. Since also, for any $f, g \in \mathcal{F}_\infty$, $\mathbb{P}_n|f^2 - g^2| \leq \mathbb{P}_n(|f-g|4F) \leq \|f-g\|_n\|4F\|_n$, we have that the covering number $N(\epsilon\|2F\|_n^2, \mathcal{F}_\infty^2, L_1(\mathbb{P}_n))$ is bounded by $N(\epsilon\|F\|_n, \mathcal{F}_\infty, L_2(\mathbb{P}_n))$. Since this last covering number is bounded by $\sup_Q N^2(\epsilon\|F\|_{Q,2}/2, \mathcal{F}, L_2(Q)) < \infty$, where the supremum is taken over all finitely discrete probability measures with $\|F\|_{Q,2} > 0$, we have by Theorem 8.14 that $\mathcal{F}_\infty^2$ is $P$-Glivenko-Cantelli. Thus $\hat{\theta}_n \xrightarrow{\mathrm{P}} 0$. This completes the proof of asymptotic equicontinuity.

The last thing we need to prove is that $\mathcal{F}$ is totally bounded in $L_2(P)$. By the result of the last paragraph, there exists a sequence of discrete probability measures $P_n$ with $\|(P_n - P)f^2\|_{\mathcal{F}_\infty} \to 0$. Fix $\epsilon > 0$ and take $n$ large enough so that $\|(P_n - P)f^2\|_{\mathcal{F}_\infty} < \epsilon^2$. Note that $N(\epsilon, \mathcal{F}, L_2(P_n))$ is finite by assumption, and, for any $f, g \in \mathcal{F}$ with $\|f-g\|_{P_n,2} < \epsilon$, $P(f-g)^2 \leq P_n(f-g)^2 + |(P_n - P)(f-g)^2| \leq 2\epsilon^2$. Thus any $\epsilon$-net in $L_2(P_n)$ is also a $\sqrt{2}\epsilon$-net in $L_2(P)$. Hence $\mathcal{F}$ is totally bounded in $L_2(P)$ since $\epsilon$ was arbitrary.□

## 8.5 Exercises

**8.5.1.** For any $\psi$ valid for defining an Orlicz norm $\|\cdot\|_\psi$, show that the space $H_\psi$ of real random variables $X$ with $\|X\|_\psi < \infty$ defines a Banach space, where we equate a random variable $X$ with zero if $X = 0$ almost surely:

(a) Show first that $\|\cdot\|_\psi$ defines a norm on $H_\psi$. Hint: Use the convexity of $\psi$ to establish that for any $X, Y \in H_\psi$ and any $c_1, c_2 > 0$,

$$\mathrm{E}\psi\left(\frac{|X+Y|}{c_1+c_2}\right) \le \frac{c_1}{c_1+c_2}\mathrm{E}\psi\left(\frac{|X|}{c_1}\right) + \frac{c_2}{c_1+c_2}\mathrm{E}\psi\left(\frac{|Y|}{c_2}\right).$$

(b) Now show that $H_\psi$ is complete. Hint: Show that for any Cauchy sequence of random variables $\{X_n\} \in H_\psi$, $\limsup_{n\to\infty} \|X_n\|_\psi < \infty$. Use Prohorov's theorem to show that every such Cauchy sequence converges to an almost surely unique element of $H_\psi$.

**8.5.2.** Show that $1 \wedge (e^u - 1)^{-1} \le 2e^{-u}$, for any $u > 0$.

**8.5.3.** For a function $\psi$ meeting the conditions of Lemma 8.2, show that there exists constants $0 < \sigma \le 1$ and $\tau > 0$ such that $\phi(x) \equiv \sigma\psi(\tau x)$ satisfies $\phi(x)\phi(y) \le \phi(cxy)$ for all $x, y \ge 1$ and $\phi(1) \le 1/2$. Show that this $\phi$ also satisfies the following

(a) For all $u > 0$, $\phi^{-1}(u) \le \psi^{-1}(u)/(\sigma\tau)$.

(b) For any random variable $X$, $\|X\|_\psi \le \|X\|_\phi/(\sigma\tau) \le \|X\|_\psi/\sigma$.

**8.5.4.** Show that for any $p \in [1, \infty)$, $\psi_p$ satisfies the conditions of Lemma 8.2 with $c = 1$.

**8.5.5.** Let $\psi$ satisfy the conditions of Lemma 8.2. Show that for any sequence of random variables $\{X_n\}$, $\|X_n\|_\psi \to 0$ implies $X_n \xrightarrow{\mathrm{P}} 0$. Hint: Show that $\liminf_{x\to\infty} \psi(x)/x > 0$, and hence $\|X_n\|_\psi \to 0$ implies $\mathrm{E}|X_n| \to 0$.

**8.5.6.** Show that if the class of functions $\mathcal{F}$ has a countable subset $\mathcal{G} \subset \mathcal{F}$ such that for each $f \in \mathcal{F}$ there exists a sequence $\{g_m\} \in \mathcal{G}$ with $g_m(x) \to f(x)$ for every $x \in \mathcal{X}$, then $\|\sum e_i f(X_i)\|_\mathcal{F} = \|\sum e_i f(X_i)\|_\mathcal{G}$ for all $(e_1, \ldots, e_n) \in \mathbb{R}^n$.

**8.5.7.** In the context of the proof of Theorem 8.15, use Hoeffding's inequality (Lemma 8.7) combined with Lemma 8.1 to show that

$$\left\| \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) \right\|_{\psi_2|X} \le \sqrt{6/n}(\mathbb{P}_n f^2)^{1/2},$$

where the meaning of the subscript $\psi_2|X$ is given in the body of the proof.

**8.5.8.** In the context of the proof of Theorem 2.3 above, show that by taking logarithms followed by square roots,

$$J_{[]}(\delta, \mathcal{F}_\infty, L_2(P)) \le \sqrt{8} J_{[]}(\delta, \mathcal{F}, L_2(P)).$$

**8.5.9.** Show, for any map $X : \Omega \mapsto \mathbb{R}$ and constants $\alpha \in [0, \infty)$ and $m \in (0, \infty)$, that

$$\mathrm{E}^*\left[|X|^\alpha 1\{|X| > m\}\right] \le m^{-1}\mathrm{E}^*\left[|X|^{\alpha+1} 1\{|X| > m\}\right].$$

## 8.6 Notes

Lemma 8.2, Corollary 8.5, and Theorems 8.15 and 8.19 correspond to Lemma 2.2.1, Corollary 2.2.8, and Theorems 2.4.3 and 2.5.2 of VW, respectively. The first inequality in Theorem 8.8 corresponds to Lemma 2.3.1 of VW. Lemma 8.3 is a modification of Lemma 2.2.10 of VW, and Theorem 8.4 is a modification and combination of both Theorem 2.2.4 and Corollary 2.2.5 of VW. The version of Hoeffding's inequality we use (Lemma 8.7) is Lemma 2.2.7 of VW, and Lemma 8.13 was inspired by Exercise 2.4.1 of VW. The proof of Theorem 2.3 follows closely van der Vaart's (1998) proof of his Theorem 19.5.

# 9
# Entropy Calculations

The focus of this chapter is on computing entropy for empirical processes. An important use of such entropy calculations is in evaluating whether a class of functions $\mathcal{F}$ is Glivenko-Cantelli and/or Donsker or neither. Several additional uses of entropy bounds will be discussed in Chapter 11. Some of these uses will be very helpful in Chapter 14 for establishing rates of convergence for M-estimators. An additional use of entropy bounds, one which will receive only limited discussion in this book, is in precisely assessing the asymptotic smoothness of certain empirical processes. Such smoothness results play a role in the asymptotic analysis of a number of statistical applications, including confidence bands for kernel density estimation (eg., Bickel and Rosenblatt, 1973) and certain hypothesis tests for multimodality (Polonik, 1995).

We begin the chapter by describing methods to evaluate uniform entropy. Provided the uniform entropy for a class $\mathcal{F}$ is not too large, $\mathcal{F}$ might be G-C or Donsker, as long as sufficient measurability holds. Since many of the techniques we will describe for building bounded uniform entropy integral (BUEI) classes (which include VC classes) closely parallel the methods for building pointwise measurable (PM) classes described in the previous chapter, we will include a discussion on joint BUEI and PM preservation towards the end of Section 9.1.2. We then present methods based on bracketing entropy. Several important results for building G-C classes from other G-C classes (*G-C preservation*), are presented next. Finally, we discuss several useful Donsker preservation results.

One can think of this chapter as a handbag of tools for establishing weak convergence properties of empirical processes. Illustrations of how to

use these tools will be given in various applications scattered throughout later chapters. To help anchor the context for these tools in practice, it might be worthwhile rereading the counting process regression example of Section 4.2.1, in the first case studies chapter of this book. In the second case studies chapter of this book (Chapter 15), we will provide additional—and more complicated—illustrations of these tools, with special emphasis on Donsker preservation techniques.

## 9.1 Uniform Entropy

We first discuss the very powerful concept of VC-classes of sets and functions. Such classes are extremely important tools in assessing and controlling bounded uniform entropy. We then discuss several useful and powerful preservation results for bounded uniform entropy integral (BUEI) classes.

### *9.1.1 VC-Classes*

In this section, we introduce Vapnik-Červonenkis (VC) classes of sets, VC-classes of functions, and several related function classes. We then present several examples of VC-classes.

Consider an arbitrary collection $\{x_1, \ldots, x_n\}$ of points in a set $\mathcal{X}$ and a collection $\mathcal{C}$ of subsets of $\mathcal{X}$. We say that $\mathcal{C}$ *picks out* a certain subset $A$ of $\{x_1, \ldots, x_n\}$ if $A = C \cap \{x_1, \ldots, x_n\}$ for some $C \in \mathcal{C}$. We say that $\mathcal{C}$ *shatters* $\{x_1, \ldots, x_n\}$ if all of the $2^n$ possible subsets of $\{x_1, \ldots, x_n\}$ are picked out by the sets in $\mathcal{C}$. The *VC-index* $V(\mathcal{C})$ of the class $\mathcal{C}$ is the smallest $n$ for which no set of size $n$ $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ is shattered by $\mathcal{C}$. If $\mathcal{C}$ shatters all sets $\{x_1, \ldots, x_n\}$ for all $n \geq 1$, we set $V(\mathcal{C}) = \infty$. Clearly, the more refined $\mathcal{C}$ is, the higher the VC-index. We say that $\mathcal{C}$ is a *VC-class* if $V(\mathcal{C}) < \infty$.

For example, let $\mathcal{X} = \mathbb{R}$ and define the collection of sets $\mathcal{C} = \{(-\infty, c] : c \in \mathbb{R}\}$. Consider any two point set $\{x_1, x_2\} \subset \mathbb{R}$ and assume, without loss of generality, that $x_1 < x_2$. It is easy to verify that $\mathcal{C}$ can pick out the null set $\{\}$ and the sets $\{x_1\}$ and $\{x_1, x_2\}$ but can't pick out $\{x_2\}$. Thus $V(\mathcal{C}) = 2$ and $\mathcal{C}$ is a VC-class. As another example, let $\mathcal{C} = \{(a, b] : -\infty \leq a < b \leq \infty\}$. The collection can shatter any two point set, but consider what happens with a three point set $\{x_1, x_2, x_3\}$. Without loss of generality, assume $x_1 < x_2 < x_3$, and note that the set $\{x_1, x_3\}$ cannot be picked out with $\mathcal{C}$. Thus $V(\mathcal{C}) = 3$ in this instance.

For any class of sets $\mathcal{C}$ and any collection $\{x_1, \ldots, x_n\} \subset \mathcal{X}$, let $\Delta_n(\mathcal{C}, x_1, \ldots, x_n)$ be the number of subsets of $\{x_1, \ldots, x_n\}$ which can be picked out by $\mathcal{C}$. A surprising combinatorial result is that if $V(\mathcal{C}) < \infty$, then $\Delta_n(\mathcal{C}, x_1, \ldots, x_n)$ can increase in $n$ no faster than $O(n^{V(\mathcal{C})-1})$. This is more precisely stated in the following lemma:

LEMMA 9.1 *For a VC-class of sets* $\mathcal{C}$,

$$\max_{x_1,\ldots,x_n\in\mathcal{X}} \Delta_n(\mathcal{C}, x_1, \ldots, x_n) \leq \sum_{j=1}^{V(\mathcal{C})-1} \left( \begin{array}{c} n \\ j \end{array} \right).$$

*Since the right-hand-side is bounded by* $V(\mathcal{C})n^{V(\mathcal{C})-1}$*, the left-hand-side grows polynomially of order at most* $O(n^{V(\mathcal{C})-1})$*.*

This is a corollary of Lemma 2.6.2 of VW, and we omit the proof.

Let $1\{\mathcal{C}\}$ denote the collection of all indicator functions of sets in the class $\mathcal{C}$. The following theorem gives a bound on the $L_r$ covering numbers of $1\{\mathcal{C}\}$:

THEOREM 9.2 *There exists a universal constant* $K < \infty$ *such that for any VC-class of sets* $\mathcal{C}$*, any* $r \geq 1$*, and any* $0 < \epsilon < 1$*,*

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

This is Theorem 2.6.4 of VW, and we omit the proof. Since $F = 1$ serves as an envelope for $1\{\mathcal{C}\}$, we have as an immediate corollary that, for $\mathcal{F} = 1\{\mathcal{C}\}$, $\sup_Q N(\epsilon\|F\|_{1,Q}, \mathcal{F}, L_1(Q)) < \infty$ and

$$J(1, \mathcal{F}, L_2) \lesssim \int_0^1 \sqrt{\log(1/\epsilon)} d\epsilon = \int_0^\infty u^{1/2} e^{-u} du \leq 1,$$

where the supremum is over all finite probability measures $Q$ with $\|F\|_{Q,2} > 0$. Thus the uniform entropy conditions required in the G-C and Donsker theorems of the previous chapter are satisfied for indicators of VC-classes of sets. Since the constant 1 serves as a universally applicable envelope function, these classes of indicator functions are therefore G-C and Donsker, provided the requisite measurability conditions hold.

For a function $f : \mathcal{X} \mapsto \mathbb{R}$, the subset of $\mathcal{X} \times \mathbb{R}$ given by $\{(x, t) : t < f(x)\}$ is the *subgraph* of $f$. A collection $\mathcal{F}$ of measurable real functions on the sample space $\mathcal{X}$ is a *VC-subgraph class* or *VC-class* (for short), if the collection of all subgraphs of functions in $\mathcal{F}$ forms a VC-class of sets (as sets in $\mathcal{X} \times \mathbb{R}$). Let $V(\mathcal{F})$ denote the VC-index of the set of subgraphs of $\mathcal{F}$. The following theorem, the proof of which is given in Section 9.5, shows that covering numbers of VC-classes of functions grow at a polynomial rate just like VC-classes of sets:

THEOREM 9.3 *There exists a universal constant* $K < \infty$ *such that, for any VC-class of measurable functions* $\mathcal{F}$ *with integrable envelope* $F$*, any* $r \geq 1$*, any probability measure* $Q$ *with* $\|F\|_{Q,r} > 0$*, and any* $0 < \epsilon < 1$*,*

$$N(\epsilon\|F\|_{Q,r}, \mathcal{F}, L_r(Q)) \leq KV(\mathcal{F})(4e)^{V(\mathcal{F})} \left(\frac{2}{\epsilon}\right)^{r(V(\mathcal{F})-1)}.$$

Thus VC-classes of functions easily satisfy the uniform entropy requirements of the G-C and Donsker theorems of the previous chapter. A related kind of function class is the *VC-hull* class. A class of measurable functions $\mathcal{G}$ is a VC-hull class if there exists a VC-class $\mathcal{F}$ such that each $f \in \mathcal{G}$ is the pointwise limit of a sequence of functions $\{f_m\}$ in the *symmetric convex hull* of $\mathcal{F}$ (denoted sconv$\mathcal{F}$). A function $f$ is in sconv$\mathcal{F}$ if $f = \sum_{i=1}^{m} \alpha_i f_i$, where the $\alpha_i$s are real numbers satisfying $\sum_{i=1}^{m} |\alpha_i| \leq 1$ and the $f_i$s are in $\mathcal{F}$. The *convex hull* of a class of functions $\mathcal{F}$, denoted conv$\mathcal{F}$, is similarly defined but with the requirement that the $\alpha_i$'s are positive. We use $\overline{\text{conv}}\mathcal{F}$ to denote pointwise closure of conv$\mathcal{F}$ and $\overline{\text{sconv}}\mathcal{F}$ to denote the pointwise closure of sconv$\mathcal{F}$. Thus the class of functions $\mathcal{F}$ is a VC-hull class if $\mathcal{F} = \overline{\text{sconv}}\mathcal{G}$ for some VC-class $\mathcal{G}$. The following theorem provides a useful relationship between the $L_2$ covering numbers of a class $\mathcal{F}$ (not necessarily a VC-class) and the $L_2$ covering numbers of $\overline{\text{conv}}\mathcal{F}$ when the covering numbers for $\mathcal{F}$ are polynomial in $1/\epsilon$:

THEOREM 9.4 *Let $Q$ be a probability measure on $(\mathcal{X}, \mathcal{A})$, and let $\mathcal{F}$ be a class of measurable functions with measurable envelope $F$, such that $QF^2 < \infty$ and, for $0 < \epsilon < 1$,*

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

*for constants $C, V < \infty$ (possibly dependent on $Q$). Then there exist a constant $K$ depending only on $V$ and $C$ such that*

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

This is Theorem 2.6.9 of VW, and we omit the proof.

It is not hard to verify that sconv$\mathcal{F}$ is a subset of the convex hull of $\mathcal{F} \cup \{-\mathcal{F}\} \cup \{0\}$, where $-\mathcal{F} \equiv \{-f : f \in \mathcal{F}\}$ (see Exercise 9.6.1 below). Since the covering numbers of $\mathcal{F} \cup \{-\mathcal{F}\} \cup \{0\}$ are at most one plus twice the covering numbers of $\mathcal{F}$, the conclusion of Theorem 9.4 also holds if $\overline{\text{conv}}\mathcal{F}$ is replaced with $\overline{\text{sconv}}\mathcal{F}$. This leads to the following easy corollary for VC-hull classes, the proof of which we save as an exercise:

COROLLARY 9.5 *For any VC-hull class $\mathcal{F}$ of measurable functions and all $0 < \epsilon < 1$,*

$$\sup_Q \log N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \leq K \left(\frac{1}{\epsilon}\right)^{2-2/V}, \quad 0 < \epsilon < 1,$$

*where the supremum is taken over all probability measures $Q$ with $\|F\|_{Q,2} > 0$, $V$ is the VC-index of the VC-subgraph class associated with $\mathcal{F}$, and the constant $K < \infty$ depends only on $V$.*

We now present several important examples and results about VC-classes of sets and both VC-subgraph and VC-hull classes of functions. The first lemma, Lemma 9.6, applies to vector spaces of functions, a frequently occurring function class in statistical applications.

LEMMA 9.6 *Let $\mathcal{F}$ be a finite-dimensional vector space of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$. Then $\mathcal{F}$ is VC-subgraph with $V(\mathcal{F}) \leq dim(\mathcal{F}) + 2$.*

**Proof.** Fix any collection $G$ of $k = \dim(\mathcal{F})+2$ points $(x_1, t_1), \ldots, (x_k, t_k)$ in $\mathcal{X} \times \mathbb{R}$. By assumption, the vectors $(f(x_1) - t_1, \ldots, f(x_k) - t_k)^T$, as $f$ ranges over $\mathcal{F}$, are restricted to a $\dim(\mathcal{F}) + 1 = k - 1$-dimensional subspace $H$ of $\mathbb{R}^k$. Any vector $0 \neq a \in \mathbb{R}^k$ orthogonal to $H$ satisfies

$$\sum_{j:a_j>0} a_j(f(x_j) - t_j) \quad = \quad \sum_{j:a_j<0} (-a_j)(f(x_j) - t_j), \tag{9.1}$$

for all $f \in \mathcal{F}$, where we define sums over empty sets to be zero. There exists such a vector $a$ with at least one strictly positive coordinate. For this $a$, the subset of $G$ of the form $\{(x_j, t_j) : a_j > 0\}$ cannot also be of the form $\{(x_j, t_j) : t_j < f(x_j)\}$ for any $f \in \mathcal{F}$, since otherwise the left side of the Equation (9.1) would be strictly positive while the right side would be nonpositive. Conclude that the subgraphs of $\mathcal{F}$ cannot pick out the subset $\{(x_j, t_j) : a_j > 0\}$. Since $G$ was arbitrary, we have just shown that the subgraphs of $\mathcal{F}$ cannot shatter any set of $k$ points in $\mathcal{X} \times \mathbb{R}$. The desired result now follows.□.

The next three lemmas, Lemmas 9.7 through 9.9, consist of useful tools for building VC-classes from other VC-classes. One of these lemmas, Lemma 9.8, is the important identity that classes of sets are VC if and only if the corresponding classes of indicator functions are VC-subgraph. The proof of Lemma 9.9 is relegated to Section 9.5.

LEMMA 9.7 *Let $\mathcal{C}$ and $\mathcal{D}$ be VC-classes of sets in a set $\mathcal{X}$, with respective VC-indices $V_\mathcal{C}$ and $V_\mathcal{D}$; and let $\mathcal{E}$ be a VC-class of sets in $\mathcal{W}$, with VC-index $V_\mathcal{E}$. Also let $\phi : \mathcal{X} \mapsto \mathcal{Y}$ and $\psi : \mathcal{Z} \mapsto \mathcal{X}$ be fixed functions. Then*

*(i) $\mathcal{C}^c \equiv \{C^c : C \in \mathcal{C}\}$ is VC with $V(\mathcal{C}^c) = V(\mathcal{C})$;*

*(ii) $\mathcal{C} \sqcap \mathcal{D} \equiv \{C \cap D : C \in \mathcal{C}, D \in \mathcal{D}\}$ is VC with index $\leq V_\mathcal{C} + V_\mathcal{D} - 1$;*

*(iii) $\mathcal{C} \sqcup \mathcal{D} \equiv \{C \cup D : C \in \mathcal{C}, D \in \mathcal{D}\}$ is VC with index $\leq V_\mathcal{C} + V_\mathcal{D} - 1$;*

*(iv) $\mathcal{D} \times \mathcal{E}$ is VC in $\mathcal{X} \times \mathcal{W}$ with VC index $\leq V_\mathcal{D} + V_\mathcal{E} - 1$;*

*(v) $\phi(\mathcal{C})$ is VC with index $V_\mathcal{C}$ if $\phi$ is one-to-one;*

*(vi) $\psi^{-1}(\mathcal{C})$ is VC with index $\leq V_\mathcal{C}$.*

**Proof.** For any $C \in \mathcal{C}$, the set $C^c$ picks out the points of a given set $x_1, \ldots, x_n$ that $C$ does not pick out. Thus if $\mathcal{C}$ shatters a given set of points, so does $\mathcal{C}^c$, and vice versa. This proves (i). 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 $O(n^{V_\mathcal{D}-1})$ further subsets. Hence $\mathcal{C} \sqcap \mathcal{D}$ can pick out at most $O(n^{V_\mathcal{C}+V_\mathcal{D}-2})$ subsets, and thus (ii) follows from the definition of a VC-class. We save it as an exercise to show that (i) and (ii) imply (iii). For (iv), note that $\mathcal{D} \times \mathcal{W}$ and $\mathcal{X} \times \mathcal{E}$ are both VC-classes with respective VC-indices $V_\mathcal{D}$ and $V_\mathcal{E}$. The desired conclusion now follows from Part (ii), since $\mathcal{D} \times \mathcal{E} = (\mathcal{D} \times \mathcal{W}) \sqcap (\mathcal{X} \times \mathcal{E})$.

For Part (v), if $\phi(\mathcal{C})$ shatters $\{y_1, \ldots, y_n\}$, then each $y_i$ must be in the range of $\phi$ and there must therefore exist $x_1, \ldots, x_n$ such that $\phi$ is a bijection between $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$. Hence $\mathcal{C}$ must shatter $x_1, \ldots, x_n$, and thus $V(\phi(\mathcal{C})) \leq V(\mathcal{C})$. Conversely, it is obvious that if $\mathcal{C}$ shatters $x_1, \ldots, x_n$, then $\phi(\mathcal{C})$ shatters $\phi(x_1), \ldots, \phi(x_n)$. Hence $V(\mathcal{C}) \leq V(\phi(\mathcal{C}))$. For (vi), if $\psi^{-1}(\mathcal{C})$ shatters $z_1, \ldots, z_n$, then all $\psi(z_i)$ must be distinct and the restriction of $\psi$ to $z_1, \ldots, z_n$ is a bijection onto its range. Thus $\mathcal{C}$ shatters $\psi(z_1), \ldots, \psi(z_n)$, and hence $V(\psi^{-1}(\mathcal{C})) \leq V(\mathcal{C})$.□

LEMMA 9.8 *For any class $\mathcal{C}$ of sets in a set $\mathcal{X}$, the class $\mathcal{F}_\mathcal{C}$ of indicator functions of sets in $\mathcal{C}$ is VC-subgraph if and only if $\mathcal{C}$ is a VC-class. Moreover, whenever at least one of $\mathcal{C}$ or $\mathcal{F}_\mathcal{C}$ is VC, the respective VC-indices are equal.*

**Proof.** Let $\mathcal{D}$ be the collection of sets of the form $\{(x,t) : t < 1\{x \in C\}\}$ for all $C \in \mathcal{C}$. Suppose that $\mathcal{D}$ is VC, and let $k = V(\mathcal{D})$. Then no set of the form $\{(x_1,0), \ldots, (x_k,0)\}$ can be shattered by $\mathcal{D}$, and hence $V(\mathcal{C}) \leq V(\mathcal{D})$. Now suppose that $\mathcal{C}$ is VC with VC-index $k$. Since for any $t < 0$, $1\{x \in C\} > t$ for all $x$ and all $C$, we have that no collection $\{(x_1,t_1), \ldots, (x_k,t_k)\}$ can be shattered by $\mathcal{D}$ if any of the $t_j$s are $< 0$. It is similarly true that no collection $\{(x_1,t_1), \ldots, (x_k,t_k)\}$ can be shattered by $\mathcal{D}$ if any of the $t_j$s are $\geq 1$, since $1\{x \in C\} > t$ is never true when $t \geq 1$. It can now be deduced that $\{(x_1,t_1), \ldots, (x_k,t_k)\}$ can only be shattered if $\{(x_1,0), \ldots, (x_k,0)\}$ can be shattered. But this can only happen if $\{x_1, \ldots, x_k\}$ can be shattered by $\mathcal{C}$. Thus $V(\mathcal{D}) \leq V(\mathcal{C})$.□

LEMMA 9.9 *Let $\mathcal{F}$ and $\mathcal{G}$ be VC-subgraph classes of functions on a set $\mathcal{X}$, with respective VC indices $V_\mathcal{F}$ and $V_\mathcal{G}$. Let $g : \mathcal{X} \mapsto \mathbb{R}$, $\phi : \mathbb{R} \mapsto \mathbb{R}$, and $\psi : \mathcal{Z} \mapsto \mathcal{X}$ be fixed functions. Then*

*(i) $\mathcal{F} \wedge \mathcal{G} \equiv \{f \wedge g : f \in \mathcal{F}, g \in \mathcal{G}\}$ is VC-subgraph with index $\leq V_\mathcal{F} + V_\mathcal{G} - 1$;*

*(ii) $\mathcal{F} \vee \mathcal{G}$ is VC with index $\leq V_\mathcal{F} + V_\mathcal{G} - 1$;*

*(iii) $\{\mathcal{F} > 0\} \equiv \{\{f > 0\} : f \in \mathcal{F}\}$ is a VC-class of sets with index $V_\mathcal{F}$;*

*(iv) $-\mathcal{F}$ is VC-subgraph with index $V_\mathcal{F}$;*

*(v)* $\mathcal{F} + g \equiv \{f + g : f \in \mathcal{F}\}$ *is VC with index* $V_{\mathcal{F}}$*;*

*(vi)* $\mathcal{F} \cdot g \equiv \{fg : f \in \mathcal{F}\}$ *is VC with index* $\leq 2V_{\mathcal{F}} - 1$*;*

*(vii)* $\mathcal{F} \circ \psi \equiv \{f(\psi) : f \in \mathcal{F}\}$ *is VC with index* $\leq V_{\mathcal{F}}$*;*

*(viii)* $\phi \circ \mathcal{F}$ *is VC with index* $\leq V_{\mathcal{F}}$ *for monotone* $\phi$*.*

The next two lemmas, Lemmas 9.10 and 9.11, refer to properties of monotone processes and classes of monotone functions. The proof of Lemma 9.11 is relegated to Section 9.5.

LEMMA 9.10 *Let* $\{X(t), t \in T\}$ *be a monotone increasing stochastic process, where* $T \subset \mathbb{R}$*. Then* $X$ *is VC-subgraph with index* $V(X) = 2$*.*

**Proof.** Let $\mathcal{X}$ be the set of all monotone increasing functions $g : T \mapsto \mathbb{R}$; and for any $s \in T$ and $x \in \mathcal{X}$, define $(s, x) \mapsto f_s(x) = x(s)$. Thus the proof is complete if we can show that the class of functions $\mathcal{F} \equiv \{f_s : s \in T\}$ is VC-subgraph with VC index 2. Now let $(x_1, t_1), (x_2, t_2)$ be any two points in $\mathcal{X} \times \mathbb{R}$. $\mathcal{F}$ shatters $(x_1, t_1), (x_2, t_2)$ if the graph $\mathcal{G}$ of $(f_s(x_1), f_s(x_2))$ in $\mathbb{R}^2$ "surrounds" the point $(t_1, t_2)$ as $s$ ranges over $T$. By surrounding a point $(a, b) \in \mathbb{R}^2$, we mean that the graph must pass through all four of the sets $\{(u, v) : u \leq a, v \leq b\}$, $\{(u, v) : u > a, v \leq b\}$, $\{(u, v) : u \leq a, v > b\}$ and $\{(u, v) : u > a, v > b\}$. By the assumed monotonicity of $x_1$ and $x_2$, the graph $\mathcal{G}$ forms a monotone curve in $\mathbb{R}^2$, and it is thus impossible for it to surround any point in $\mathbb{R}^2$. Thus $(x_1, t_1), (x_2, t_2)$ cannot be shattered by $\mathcal{F}$, and the desired result follows.□

LEMMA 9.11 *The set* $\mathcal{F}$ *of all monotone functions* $f : \mathbb{R} \mapsto [0, 1]$ *satisfies*

$$\sup_Q \log N(\epsilon, \mathcal{F}, L_2(Q)) \leq \frac{K}{\epsilon}, \quad 0 < \epsilon < 1,$$

*where the supremum is taken over all probability measures* $Q$*, and the constant* $K < \infty$ *is universal.*

The final lemma of this section, Lemma 9.12 below, addresses the claim raised in Section 4.1 that the class of functions $\mathcal{F} \equiv \{1\{Y - b'Z \leq v\} : b \in \mathbb{R}^k, v \in \mathbb{R}\}$ is Donsker. Because the indicator functions in $\mathcal{F}$ are a subset of the indicator functions for half-spaces in $\mathbb{R}^{k+1}$, Part (i) of the lemma implies that $\mathcal{F}$ is VC with index $k + 3$. Since Lemma 8.12 from Chapter 8 verifies that $\mathcal{F}$, $\mathcal{F}_\delta$ and $\mathcal{F}_\infty^2$ are all $P$-measurable, for any probability measure $P$, Theorem 9.3 combined with Theorem 8.19 and the fact that indicator functions are bounded, establishes that $\mathcal{F}$ is $P$-Donsker for any $P$. Lemma 9.12 also gives a related result on closed balls in $\mathbb{R}^d$. In the lemma, $\langle a, b \rangle$ denotes the Euclidean inner product.

LEMMA 9.12 *The following are true:*

*(i) The collection of all* half-spaces *in* $\mathbb{R}^d$*, consisting of the sets* $\{x \in \mathbb{R}^d : \langle x, u\rangle \leq c\}$ *with* $u$ *ranging over* $\mathbb{R}^d$ *and* $c$ *ranging over* $\mathbb{R}$*, is VC with index* $d+2$*.*

*(ii) The collection of all closed balls in* $\mathbb{R}^d$ *is VC with index* $\leq d+3$*.*

**Proof.** The class $\mathcal{A}^+$ of sets $\{x : \langle x, u\rangle \leq c\}$, with $u$ ranging over $\mathbb{R}^d$ and $c$ ranging over $(0, \infty)$, is equivalent to the class of sets $\{x : \langle x, u\rangle - 1 \leq 0\}$ with $u$ ranging over $\mathbb{R}^d$. In this last class, since $\langle x, u\rangle$ spans a $d$-dimensional vector space, Lemma 9.6 and Part (v) of Lemma 9.9 yield that the class of functions spanned by $\langle x, u\rangle - 1$ is VC with index $d+2$. Part (iii) of Lemma 9.9 combined with Part (i) of Lemma 9.7 now yields that the class $\mathcal{A}^+$ is VC with index $d+2$. Similar arguments verify that both the class $\mathcal{A}^-$, with $c$ restricted to $(-\infty, 0)$, and the class $\mathcal{A}^0$, with $c = 0$, are VC with index $d+2$. It is easy to verify that the union of finite VC classes has VC index equal to the maximum of the respective VC indices. This concludes the proof of (i).

Closed balls in $\mathbb{R}^d$ are sets of the form $\{x : \langle x, x\rangle - 2\langle x, u\rangle + \langle u, u\rangle \leq c\}$, where $u$ ranges over $\mathbb{R}^d$ and $c$ ranges over $[0, \infty)$. It is straightforward to check that the class $\mathcal{G}$ all functions of the form $x \mapsto -2\langle x, u\rangle + \langle u, u\rangle - c$ are contained in a $d+1$ dimensional vector space, and thus $\mathcal{G}$ is VC with index $\leq d+3$. Combining this with Part (v) of Lemma 9.9 yields that the class $\mathcal{F} = \mathcal{G} + \langle x, x\rangle$ is also VC with index $d+3$. Now the desired conclusion follows from Part (iii) if Lemma 9.9 combined with Part (i) of Lemma 9.7.□

### *9.1.2 BUEI Classes*

Recall for a class of measurable functions $\mathcal{F}$, with envelope $F$, the uniform entropy integral

$$J(\delta, \mathcal{F}, L_2) \equiv \int_0^\delta \sqrt{\sup_Q \log N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))}\, d\epsilon,$$

where the supremum is taken over all finitely discrete probability measures $Q$ with $\|F\|_{Q,2} > 0$. Note the dependence on choice of envelope $F$. This is crucial since there are many random functions which can serve as an envelope. For example, if $F$ is an envelope, then so is $F+1$ and $2F$. One must allow that different envelopes may be needed in different settings. We say that the class $\mathcal{F}$ has *bounded uniform entropy integral* (BUEI) with envelope $F$—or is *BUEI* with envelope $F$—if $J(1, \mathcal{F}, L_2) < \infty$ for that particular choice of envelope.

Theorem 9.3 tells us that a VC-class $\mathcal{F}$ is automatically BUEI with any envelope. We leave it as an exercise to show that if $\mathcal{F}$ and $\mathcal{G}$ are BUEI with respective envelopes $F$ and $G$, then $\mathcal{F} \sqcup \mathcal{G}$ is BUEI with envelope $F \vee G$. The following lemma, which is closely related to an important Donsker

preservation theorem in Section 9.4 below, is also useful for building BUEI classes from other BUEI classes:

LEMMA 9.13 *Let $\mathcal{F}_1, \ldots, \mathcal{F}_k$ be BUEI classes with respective envelopes $F_1, \ldots, F_k$, and let $\phi : \mathbb{R}^k \mapsto \mathbb{R}$ satisfy*

$$|\phi \circ f(x) - \phi \circ g(x)|^2 \leq c^2 \sum_{j=1}^{k} (f_j(x) - g_j(x))^2, \tag{9.2}$$

*for every $f, g \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_k$ and $x$ for a constant $0 < c < \infty$. Then the class $\phi \circ (\mathcal{F}_1, \ldots, \mathcal{F}_k)$ is BUEI with envelope $H \equiv |\phi(f_0)| + c \sum_{j=1}^{k} (|f_{0j}| + F_j)$, where $f_0 \equiv (f_{01}, \ldots, f_{0k})$ is any function in $\mathcal{F}_1 \times \cdots \times \mathcal{F}_k$, and where $\phi \circ (\mathcal{F}_1, \ldots, \mathcal{F}_k)$ is as defined in Lemma 8.10.*

**Proof.** Fix $\epsilon > 0$ and a finitely discrete probability measure $Q$, and let $f, g \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_k$ satisfy $\|f_j - g_j\|_{Q,2} \leq \epsilon \|F_j\|_{Q,2}$ for $1 \leq j \leq k$. Now (9.2) implies that

$$\begin{aligned} \|\phi \circ f - \phi \circ g\|_{Q,2} &\leq c \sqrt{\sum_{j=1}^{k} \|f_j - g_j\|_{Q,2}^2} \\ &\leq \epsilon c \sum_{j=1}^{k} \|F_j\|_{Q,2} \\ &\leq \epsilon \|H\|_{Q,2}. \end{aligned}$$

Hence

$$N(\epsilon \|H\|_{Q,2}, \phi \circ (\mathcal{F}_1, \ldots, \mathcal{F}_k), L_2(Q)) \leq \prod_{j=1}^{k} N(\epsilon \|F_j\|_{Q,2}, \mathcal{F}_j, L_2(Q)),$$

and the desired result follows since $\epsilon$ and $Q$ were arbitrary.□

Some useful consequences of Lemma 9.13 are given in the following lemma:

LEMMA 9.14 *Let $\mathcal{F}$ and $\mathcal{G}$ be BUEI with respective envelopes $F$ and $G$, and let $\phi : \mathbb{R} \mapsto \mathbb{R}$ be a Lipschitz continuous function with Lipschitz constant $0 < c < \infty$. Then*

*(i) $\mathcal{F} \wedge \mathcal{G}$ is BUEI with envelope $F + G$;*

*(ii) $\mathcal{F} \vee \mathcal{G}$ is BUEI with envelope $F + G$;*

*(iii) $\mathcal{F} + \mathcal{G}$ is BUEI with envelope $F + G$;*

*(iv) $\phi(\mathcal{F})$ is BUEI with envelope $|\phi(f_0)| + c(|f_0| + F)$, provided $f_0 \in \mathcal{F}$.*

The proof, which we omit, is straightforward.

As mentioned earlier, Lemma 9.13 is very similar to a Donsker preservation result we will present later in this chapter. In fact, most of the BUEI preservation results we give in this section have parallel Donsker preservation properties. An important exception, and one which is perhaps the primary justification for the use of BUEI preservation techniques, applies to products of Donsker classes. As verified in the following theorem, the product of two BUEI classes is BUEI, whether or not the two classes involved are bounded (compare with Corollary 9.15 below):

THEOREM 9.15 *Let $\mathcal{F}$ and $\mathcal{G}$ be BUEI classes with respective envelopes $F$ and $G$. Then $\mathcal{F} \cdot \mathcal{G} \equiv \{fg : f \in \mathcal{F}, g \in \mathcal{G}\}$ is BUEI with envelope $FG$.*

**Proof.** Fix $\epsilon > 0$ and a finitely discrete probability measure $\tilde{Q}$ with $\|FG\|_{\tilde{Q},2} > 0$, and let $dQ^* \equiv G^2 d\tilde{Q}/\|G\|^2_{\tilde{Q},2}$. Clearly, $Q^*$ is a finitely discrete probability measure with $\|F\|_{Q^*,2} > 0$. Let $f_1, f_2 \in \mathcal{F}$ satisfy $\|f_1 - f_2\|_{Q^*,2} \leq \epsilon \|F\|_{Q^*,2}$. Then

$$\epsilon \geq \frac{\|f_1 - f_2\|_{Q^*,2}}{\|F\|_{Q^*,2}} = \frac{\|(f_1 - f_2)G\|_{\tilde{Q},2}}{\|FG\|_{\tilde{Q},2}},$$

and thus, if we let $\mathcal{F} \cdot G \equiv \{fG : f \in \mathcal{F}\}$,

$$\begin{aligned} N(\epsilon \|FG\|_{\tilde{Q},2}, \mathcal{F} \cdot G, L_2(\tilde{Q})) &\leq N(\epsilon \|F\|_{Q^*,2}, \mathcal{F}, L_2(Q^*)) \\ &\leq \sup_Q N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)), \end{aligned} \tag{9.3}$$

where the supremum is taken over all finitely discrete probability measures $Q$ for which $\|F\|_{Q,2} > 0$. Since the right-hand-side of (9.3) does not depend on $\tilde{Q}$, and since $\tilde{Q}$ satisfies $\|FG\|_{\tilde{Q},2} > 0$ but is otherwise arbitrary, we have that

$$\sup_Q N(\epsilon \|FG\|_{Q,2}, \mathcal{F} \cdot G, L_2(Q)) \leq \sup_Q N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)),$$

where the supremums are taken over all finitely discrete probability measures $Q$ but with the left side taken over the subset for which $\|FG\|_{Q,2} > 0$ while the right side is taken over the subset for which $\|F\|_{Q,2} > 0$.

We can similarly show that the uniform entropy numbers for the class $\mathcal{G} \cdot F$ with envelope $FG$ is bounded by the uniform entropy numbers for $\mathcal{G}$ with envelope $G$. Since $|f_1 g_1 - f_2 g_2| \leq |f_1 - f_2|G + |g_1 - g_2|F$ for all $f_1, f_2 \in \mathcal{F}$ and $g_1, g_2 \in \mathcal{G}$, the forgoing results imply that

$$\begin{aligned} \sup_Q N(\epsilon \|FG\|_{Q,2}, \mathcal{F} \cdot \mathcal{G}, L_2(Q)) &\leq \sup_Q N(\epsilon \|F\|_{Q,2}/2, \mathcal{F}, L_2(Q)) \\ &\quad \times \sup_Q N(\epsilon \|G\|_{Q,2}/2, \mathcal{G}, L_2(Q)), \end{aligned}$$

where the supremums are all taken over the appropriate subsets of all finitely discrete probability measures. After taking logs, square roots, and then integrating both sides with respect to $\epsilon$, the desired conclusion follows.□

In order for BUEI results to be useful for obtaining Donsker results, it is necessary that sufficient measurability be established so that Theorem 8.19 can be used. As shown in Proposition 8.11 and the comments following Theorem 8.19, pointwise measurability (PM) is sufficient measurability for this purpose. Since there are significant similarities between PM preservation and BUEI preservation results, one can construct useful joint PM and BUEI preservation results. Here is one such result:

LEMMA 9.16 *Let the classes $\mathcal{F}_1, \ldots, \mathcal{F}_k$ be both BUEI and PM with respective envelopes $F_1, \ldots, F_k$, and let $\phi : \mathbb{R}^k \mapsto \mathbb{R}$ satisfy (9.2) for every $f, g \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_k$ and $x$ for a constant $0 < c < \infty$. Then the class $\phi \circ (\mathcal{F}_1, \ldots, \mathcal{F}_k)$ is both BUEI and PM with envelope $H \equiv |\phi(f_0)| + c\sum_{j=1}^{k}(|f_{0j}| + F_j)$, where $f_0$ is any function in $\mathcal{F}_1 \times \cdots \times \mathcal{F}_k$.*

**Proof.** Since a function satisfying (9.2) as specified is also continuous, the desired result is a direct consequence of Lemmas 8.10 and 9.13.□

The following lemma contains some additional joint preservation results:

LEMMA 9.17 *Let the classes $\mathcal{F}$ and $\mathcal{G}$ be both BUEI and PM with respective envelopes $F$ and $G$, and let $\phi : \mathbb{R} \mapsto \mathbb{R}$ be a Lipschitz continuous function with Lipschitz constant $0 < c < \infty$. Then*

*(i) $\mathcal{F} \cup \mathcal{G}$ is both BUEI and PM with envelope $F \vee G$;*

*(ii) $\mathcal{F} \wedge \mathcal{G}$ is both BUEI and PM with envelope $F + G$;*

*(iii) $\mathcal{F} \vee \mathcal{G}$ is both BUEI and PM with envelope $F + G$;*

*(iv) $\mathcal{F} + \mathcal{G}$ is both BUEI and PM with envelope $F + G$;*

*(v) $\mathcal{F} \cdot \mathcal{G}$ is both BUEI and PM with envelope $FG$;*

*(vi) $\phi(\mathcal{F})$ is both BUEI and PM with envelope $|\phi(f_0)| + c(|f_0| + F)$, where $f_0 \in \mathcal{F}$.*

**Proof.** Verifying (i) is straightforward. Results (ii), (iii), (iv) and (vi) are consequences of Lemma 9.16. Result (v) is a consequence of Lemma 8.10 and Theorem 9.15.□

If a class of measurable functions $\mathcal{F}$ is both BUEI and PM with envelope $F$, then Theorem 8.19 implies that $\mathcal{F}$ is $P$-Donsker whenever $P^*F^2 < \infty$. Note that we have somehow avoided discussing preservation for subsets of classes. This is because it is unclear whether a subset of a PM class $\mathcal{F}$ is itself a PM class. The difficulty is that while $\mathcal{F}$ may have a countable dense subset $\mathcal{G}$ (dense in terms of pointwise convergence), it is unclear whether any arbitrary subset $\mathcal{H} \subset \mathcal{F}$ also has a suitable countable dense subset. An easy way around this problem is to use various preservation results to

establish that $\mathcal{F}$ is $P$-Donsker, and then it follows directly that any $\mathcal{H} \subset \mathcal{F}$ is also $P$-Donsker by the definition of weak convergence. We will explore several additional preservation results as well as several practical examples later in this chapter and in the case studies of Chapter 15.

## 9.2 Bracketing Entropy

We now present several useful bracketing entropy results for certain function classes as well as a few preservation results. We first mention that bracketing numbers are in general larger than covering numbers, as verified in the following lemma:

LEMMA 9.18 *Let $\mathcal{F}$ be any class of real function on $\mathcal{X}$ and $\|\cdot\|$ any norm on $\mathcal{F}$. Then*

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

*for all $\epsilon > 0$.*

**Proof.** Fix $\epsilon > 0$, and let $\mathcal{B}$ be collection of $\epsilon$-brackets that covers $\mathcal{F}$. From each bracket $B \in \mathcal{B}$, take a function $g(B) \in B \cap \mathcal{F}$ to form a finite collection of functions $\mathcal{G} \subset \mathcal{F}$ of the same cardinality as $\mathcal{B}$ consisting of one function from each bracket in $\mathcal{B}$. Now every $f \in \mathcal{F}$ lies in a bracket $B \in \mathcal{B}$ such that $\|f - g(B)\| \leq \epsilon$ by the definition of an $\epsilon$-bracket. Thus $\mathcal{G}$ is an $\epsilon$ cover of $\mathcal{F}$ of the same cardinality as $\mathcal{B}$. The desired conclusion now follows.□

The first substantive bracketing entropy result we present considers classes of smooth functions on a bounded set $\mathcal{X} \subset \mathbb{R}^d$. For any vector $k = (k_1, \ldots, k_d)$ of nonnegative integers define the differential operator $D^k \equiv \partial^{|k|}/(\partial x_1^{k_1}, \ldots, \partial x_d^{k_d})$, where $|k| \equiv k_1 + \cdots + k_d$. As defined previously, let $\lfloor x \rfloor$ be the largest integer $j \leq x$, for any $x \in \mathbb{R}$. For any function $f : \mathcal{X} \mapsto \mathbb{R}$ and $\alpha > 0$, define the norm

$$\|f\|_\alpha \equiv \max_{k:|k|\leq\lfloor\alpha\rfloor} \sup_x |D^k f(x)| + \max_{k:|k|=\lfloor\alpha\rfloor} \sup_{x,y} \frac{|D^k f(x) - D^k f(y)|}{\|x - y\|^{\alpha - \lfloor\alpha\rfloor}},$$

where the suprema are taken over $x \neq y$ in the interior of $\mathcal{X}$. When $k = 0$, we set $D^k f = f$. Now let $C_M^\alpha(\mathcal{X})$ be the set of all continuous functions $f : \mathcal{X} \mapsto \mathbb{R}$ with $\|f\|_\alpha \leq M$. Recall that for a set $A$ in a metric space, $\operatorname{diam} A = \sup_{x,y\in A} d(x, y)$. We have the following theorem:

THEOREM 9.19 *Let $\mathcal{X} \subset \mathbb{R}^d$ be bounded and convex with nonempty interior. There exists a constant $K < \infty$ depending only on $\alpha$,* $\operatorname{diam} \mathcal{X}$*, and $d$ such that*

$$\log N_{[]}(\epsilon, C_1^\alpha(\mathcal{X}), L_r(Q)) \leq K \left(\frac{1}{\epsilon}\right)^{d/\alpha},$$

*for every $r \geq 1$, $\epsilon > 0$, and any probability measure $Q$ on $\mathbb{R}^d$.*

This is Corollary 2.7.2 of VW, and we omit the proof.

We now present a generalization of Theorem 9.19 which permits $\mathcal{X}$ to be unbounded:

THEOREM 9.20 *Let $\mathbb{R}^d = \bigcup_{j=1}^{\infty} I_j$ be a partition of $\mathbb{R}^d$ into bounded, convex sets with nonempty interior, and let $\mathcal{F}$ be a class of measurable functions $f : \mathbb{R}^d \mapsto \mathbb{R}$ such that the restrictions $\mathcal{F}_{|I_j}$ belong to $C^{\alpha}_{M_j}$ for all $j$. Then, for every $V \geq d/\alpha$, there exists a constant $K$ depending only on $\alpha$, $r$, and $d$, such that*

$$\log N_{[]}(\epsilon, \mathcal{F}, L_r(Q)) \leq K \left(\frac{1}{\epsilon}\right)^V \left[\sum_{j=1}^{\infty} \lambda(I_j^1)^{\frac{r}{V+r}} M_j^{\frac{Vr}{V+r}} Q(I_j)^{\frac{V}{V+r}}\right]^{\frac{V+r}{r}},$$

*for every $\epsilon > 0$ and probability measure $Q$, where, for a set $A \subset \mathbb{R}^d$, $A^1 \equiv \{x : \|x - A\| < 1\}$.*

This is Corollary 2.7.4 of VW, and we omit the proof.

We now consider several results for Lipschitz and Sobolev function classes. We first present the results for covering numbers based on the uniform norm and then present the relationship to bracketing entropy.

THEOREM 9.21 *For a compact, convex subset $C \subset \mathbb{R}^d$, let $\mathcal{F}$ be the class of all convex functions $f : C \mapsto [0,1]$ with $|f(x) - f(y)| \leq L\|x - y\|$ for every $x, y$. For some integer $m \geq 1$, let $\mathcal{G}$ be the class of all functions $g : [0,1] \mapsto [0,1]$ with $\int_0^1 [g^{(m)}(x)]^2 dx \leq 1$, where superscript $(m)$ denotes the $m$'th derivative. Then*

$$\begin{aligned} \log N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) &\leq K(1+L)^{d/2} \left(\frac{1}{\epsilon}\right)^{d/2}, \quad \textit{and} \\ \log N(\epsilon, \mathcal{G}, \|\cdot\|_\infty) &\leq M \left(\frac{1}{\epsilon}\right)^{1/m}, \end{aligned}$$

*where $\|\cdot\|_\infty$ is the uniform norm and the constant $K < \infty$ depends only on $d$ and $C$ and the constant $M$ depends only on $m$.*

The first displayed result is Corollary 2.7.10 of VW, while the second displayed result is Theorem 2.4 of van de Geer (2000). We omit the proofs.

The following lemma shows how Theorem 9.21 applies to bracketing entropy:

LEMMA 9.22 *For any norm $\|\cdot\|$ dominated by $\|\cdot\|_\infty$ and any class of functions $\mathcal{F}$,*

$$\log N_{[]}(2\epsilon, \mathcal{F}, \|\cdot\|) \leq \log N(\epsilon, \mathcal{F}, \|\cdot\|_\infty),$$

*for all $\epsilon > 0$.*

**Proof.** Let $f_1, \ldots, f_m$ be a uniform $\epsilon$-cover of $\mathcal{F}$. Since the $2\epsilon$-brackets $[f_i - \epsilon, f_i + \epsilon]$ now cover $\mathcal{F}$, the result follows.□

We now present a second Lipschitz continuity result which is in fact a generalization of Lemma 9.22. The result applies to function classes of the form $\mathcal{F} = \{f_t : t \in T\}$, where

$$|f_s(x) - f_t(x)| \quad \leq \quad d(s,t)F(x) \tag{9.4}$$

for some metric $d$ on $T$, some real function $F$ on the sample space $\mathcal{X}$, and for all $x \in \mathcal{X}$. This special Lipschitz structure arises in a number of settings, including parametric Z- and M- estimation. For example, consider the least absolute deviation regression setting of Section 2.2.6, under the assumption that the random covariate $U$ and regression parameter $\theta$ are constrained to known compact subsets $\mathcal{U}, \Theta \subset \mathbb{R}^p$. Recall that, in this setting, the outcome given $U$ is modeled as $Y = \theta' U + e$, where the residual error $e$ has median zero. Estimation of the true parameter value $\theta_0$ is accomplished by minimizing $\theta \mapsto \mathbb{P}_n m_\theta$, where $m_\theta(X) \equiv |e - (\theta - \theta_0)'U| - |e|$, $X \equiv (Y, U)$ and $e = Y - \theta_0' U$. From (2.20) in Section 2.2.6, we know that the class $\mathcal{F} = \{m_\theta : \theta \in \Theta\}$ satisfies (9.4) with $T = \Theta$, $d(s,t) = \|s - t\|$ and $F(x) = \|u\|$, where $x = (y, u)$ is a realization of $X$.

The following theorem shows that the bracketing numbers for a general $\mathcal{F}$ satisfying (9.4) are bounded by the covering numbers for the associated index set $T$.

THEOREM 9.23 *Suppose the class of functions $\mathcal{F} = \{f_t : t \in T\}$ satisfies (9.4) for every $s, t \in T$ and some fixed function $F$. Then, for any norm $\|\cdot\|$,*

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

**Proof.** Note that for any $\epsilon$-net $t_1, \ldots, t_k$ that covers $T$ with respect to $d$, the brackets $[f_{t_j} - \epsilon F, f_{t_j} + \epsilon F]$ cover $\mathcal{F}$. Since these brackets are all of size $2\epsilon\|F\|$, the proof is complete.□

Note that when $\|\cdot\|$ is any norm dominated by $\|\cdot\|_\infty$, Theorem 9.23 simplifies to Lemma 9.22 when $T = \mathcal{F}$ and $d = \|\cdot\|_\infty$ (and thus automatically $F = 1$).

We move now from continuous functions to monotone functions. As was done in Lemma 9.11 above for uniform entropy, we can study bracketing entropy of the class of all monotone functions mapping into $[0, 1]$:

THEOREM 9.24 *For each integer $r \geq 1$, there exists a constant $K < \infty$ such that the class $\mathcal{F}$ of monotone functions $f : \mathbb{R} \mapsto [0, 1]$ satisfies*

$$\log N_{[]}(\epsilon, \mathcal{F}, L_r(Q)) \leq \frac{K}{\epsilon},$$

*for all $\epsilon > 0$ and every probability measure $Q$.*

The lengthy proof, which we omit, is given in Chapter 2.7 of VW.

We now briefly discuss preservation results. Unfortunately, it appears that there are not as many useful preservation results for bracketing entropy as there are for uniform entropy, but the following lemma contains two such results which are easily verified:

LEMMA 9.25 *Let $\mathcal{F}$ and $\mathcal{G}$ be classes of measurable function. Then for any probability measure $Q$ and any $1 \leq r \leq \infty$,*

*(i)* $N_{[]}(2\epsilon, \mathcal{F}+\mathcal{G}, L_r(Q)) \leq N_{[]}(\epsilon, \mathcal{F}, L_r(Q))N_{[]}(\epsilon, \mathcal{G}, L_r(Q))$*;*

*(ii) Provided $\mathcal{F}$ and $\mathcal{G}$ are bounded by 1,*

$$N_{[]}(2\epsilon, \mathcal{F}\cdot\mathcal{G}, L_r(Q)) \leq N_{[]}(\epsilon, \mathcal{F}, L_r(Q))N_{[]}(\epsilon, \mathcal{G}, L_r(Q)).$$

The straightforward proof is saved as an exercise.

## 9.3 Glivenko-Cantelli Preservation

In this section, we discuss methods which are useful for building up Glivenko-Cantelli (G-C) classes from other G-C classes. Such results can be useful for establishing consistency for Z- and M- estimators and their bootstrapped versions. It is clear from the definition of $P$-G-C classes, that if $\mathcal{F}$ and $\mathcal{G}$ are $P$-G-C, then $\mathcal{F} \cup \mathcal{G}$ and any subset thereof is also $P$-G-C. The purpose of the remainder of this section is to discuss more substantive preservation results. The main tool for this is the following theorem, which is a minor modification of Theorem 3 of van der Vaart and Wellner (2000) and which we give without proof:

THEOREM 9.26 *Suppose that $\mathcal{F}_1, \ldots, \mathcal{F}_k$ are strong $P$-G-C classes of functions with $\max_{1\leq j\leq k} \|P\|_{\mathcal{F}_j} < \infty$, and that $\phi : \mathbb{R}^k \mapsto \mathbb{R}$ is continuous. Then the class $\mathcal{H} \equiv \phi(\mathcal{F}_1, \ldots, \mathcal{F}_k)$ is strong $P$-G-C provided it has an integrable envelope.*

The following corollary lists some obvious consequences of this theorem:

COROLLARY 9.27 *Let $\mathcal{F}$ and $\mathcal{G}$ be $P$-G-C classes with respective integrable envelopes $F$ and $G$. Then the following are true:*

*(i) $\mathcal{F}+\mathcal{G}$ is $P$-G-C.*

*(ii) $\mathcal{F}\cdot\mathcal{G}$ is $P$-G-C provided $P[FG] < \infty$.*

*(iii) Let $R$ be the union of the ranges of functions in $\mathcal{F}$, and let $\psi : \overline{R} \mapsto \mathbb{R}$ be continuous. Then $\psi(\mathcal{F})$ is $P$-G-C provided it has an integrable envelope.*

**Proof.** The statement (i) is obvious. Since $(x, y) \mapsto xy$ is continuous in $\mathbb{R}^2$, statement (ii) follows from Theorem 9.26. Statement (iii) also follows from the theorem since $\psi$ has a continuous extension to $\mathbb{R}$, $\tilde{\psi}$, such that $\|P\tilde{\psi}(f)\|_{\mathcal{F}} = \|P\psi(f)\|_{\mathcal{F}}$.□

It is interesting to note that the "preservation of products" result in Part (ii) of the above corollary does not hold in general for Donsker classes (although, as was shown in Section 9.1.2, it does hold for BUEI classes). This preservation result for G-C classes can be useful in formulating master theorems for bootstrapped Z- and M- estimators. Consider, for example, verifying the validity of the bootstrap for a parametric Z-estimator $\hat{\theta}_n$ which is a zero of $\theta \mapsto \mathbb{P}_n\psi_\theta$, for $\theta \in \Theta$, where $\psi_\theta$ is a suitable random function. Let $\Psi(\theta) = P\psi_\theta$, where we assume that for any sequence $\{\theta_n\} \in \Theta$, $\Psi(\theta_n) \to 0$ implies $\theta_n \to \theta_0 \in \Theta$ (i.e., the parameter is identifiable). Usually, to obtain consistency, it is reasonable to assume that the class $\{\psi_\theta, \theta \in \Theta)$ is $P$-G-C. Clearly, this condition is sufficient to ensure that $\hat{\theta}_n \overset{\text{as*}}{\to} \theta_0$.

Now, under a few additional assumptions, the Z-estimator master theorem, Theorem 2.11 can be applied, to obtain asymptotic normality of $\sqrt{n}(\hat{\theta}_n - \theta_0)$. In Section 2.2.5, we made the claim that if $\Psi$ is appropriately differentiable and the parameter is identifiable (as defined in the previous paragraph), sufficient additional conditions for this asymptotic normality to hold and for the bootstrap to be valid are that the $\{\psi_\theta : \theta \in \Theta\}$ is strong $P$-G-C with $\sup_{\theta\in\Theta} P|\psi_\theta| < \infty$, that $\{\psi_\theta : \theta \in \Theta, \|\theta - \theta_0\| \le \delta\}$ is $P$-Donsker for some $\delta > 0$, and that $P\|\psi_\theta - \psi_{\theta_0}\|^2 \to 0$ as $\theta \to \theta_0$. As we will see in Chapter 13, where we present the arguments for this result in detail, an important step in the proof of bootstrap validity is to show that the bootstrap estimate $\hat{\theta}_n^\circ$ is unconditionally consistent for $\theta_0$. If we use a weighted bootstrap with i.i.d. non-negative weights $\xi_1, \ldots, \xi_n$, which are independent of the data and which satisfy $\mathrm{E}\xi_1 = 1$, then result (ii) from the above corollary tells us that $\mathcal{F} \equiv \{\xi\psi_\theta : \theta \in \Theta\}$ is $P$-G-C. This follows since both classes of functions $\{\xi\}$ (a trivial class with one member) and $\{\psi_\theta : \theta \in \Theta\}$ are $P$-G-C and since the product class $\mathcal{F}$ has an integral envelope by Lemma 8.13. Note here that we are tacitly augmenting $P$ to be the product probability measure of both the data and the independent bootstrap weights. We will expand on these ideas in Section 10.3 of the next chapter for the special case where $\Theta \subset \mathbb{R}^p$ and in Chapter 13 for the more general case.

Another result that can be useful for inference is the following lemma on covariance estimation. We mentioned this result in the first paragraph of Section 2.2.3 in the context of conducting uniform inference for $Pf$ as $f$ ranges over a class of functions $\mathcal{F}$. The lemma answers the question of when the limiting covariance of $\mathbb{G}_n$, indexed by $\mathcal{F}$, can be consistently estimated. Recall that this covariance is $\sigma(f, g) \equiv Pfg - PfPg$, and its estimator is $\hat{\sigma}(f, g) \equiv \mathbb{P}_n fg - \mathbb{P}_n f\mathbb{P}_n g$. Although knowledge of this covariance matrix is

usually not sufficient in itself to obtain inference on $\{Pf : f \in \mathcal{F}\}$, it still provides useful information.

LEMMA 9.28 *Let $\mathcal{F}$ be Donsker. Then $\|\hat{\sigma}(f,g) - \sigma(f,g)\|_{\mathcal{F}.\mathcal{F}} \overset{\text{as*}}{\to} 0$ if and only if $P^*\|f - Pf\|_{\mathcal{F}}^2 < \infty$.*

**Proof.** Note that since $\mathcal{F}$ is Donsker, $\mathcal{F}$ is also G-C. Hence $\dot{\mathcal{F}} \equiv \{\dot{f} : f \in \mathcal{F}\}$ is G-C, where for any $f \in \mathcal{F}$, $\dot{f} = f - Pf$. Now we first assume that $P^*\|f - Pf\|_{\mathcal{F}}^2 < \infty$. By Theorem 9.26, $\dot{\mathcal{F}} \cdot \dot{\mathcal{F}}$ is also G-C. Uniform consistency of $\hat{\sigma}$ now follows since, for any $f, g \in \mathcal{F}$, $\hat{\sigma}(f,g) - \sigma(f,g) = (\mathbb{P}_n - P)\dot{f}\dot{g} - \mathbb{P}_n\dot{f}\mathbb{P}_n\dot{g}$. Assume next that $\|\hat{\sigma}(f,g) - \sigma(f,g)\|_{\mathcal{F}.\mathcal{F}} \overset{\text{as*}}{\to} 0$. This implies that $\dot{\mathcal{F}} \cdot \dot{\mathcal{F}}$ is G-C. Now Lemma 8.13 implies that $P^*\|f - Pf\|_{\mathcal{F}}^2 = P^*\|fg\|_{\dot{\mathcal{F}}.\dot{\mathcal{F}}} < \infty$.□

We close this section with the following theorem that provides several interesting necessary and sufficient conditions for $\mathcal{F}$ to be strong G-C:

THEOREM 9.29 *Let $\mathcal{F}$ be a class of measurable functions. Then the following are equivalent:*

*(i) $\mathcal{F}$ is strong $P$-G-C;*

*(ii) $\mathrm{E}^*\|\mathbb{P}_n - P\|_{\mathcal{F}} \to 0$ and $\mathrm{E}^*\|f - Pf\|_{\mathcal{F}} < \infty$;*

*(iii) $\|\mathbb{P}_n - P\|_{\mathcal{F}} \overset{\text{P}}{\to} 0$ and $\mathrm{E}^*\|f - Pf\|_{\mathcal{F}} < \infty$.*

**Proof.** Since $\mathbb{P}_n - P$ does not change when the class $\mathcal{F}$ is replaced by $\{f - Pf : f \in \mathcal{F}\}$, we will assume hereafter that $\|P\|_{\mathcal{F}} = 0$ without loss of generality.

(i)⇒(ii): That (i) implies $\mathrm{E}^*\|f\|_{\mathcal{F}} < \infty$ follows from Lemma 8.13. Fix $0 < M < \infty$, and note that

$$\begin{aligned} \mathrm{E}^*\|\mathbb{P}_n - P\|_{\mathcal{F}} \quad &\leq \quad \mathrm{E}^*\left\|(\mathbb{P}_n - P)f \times 1\{F \leq M\}\right\|_{\mathcal{F}} \\ &\quad +2\mathrm{E}^*\left[F \times 1\{F > M\}\right]. \end{aligned} \tag{9.5}$$

By Assertion (ii) of Corollary 9.27, $\mathcal{F} \cdot 1\{F \leq M\}$ is strong $P$-G-C, and thus the first term on the right of (9.5) $\to 0$ by the bounded convergence theorem. Since the second term on the right of (9.5) can be made arbitrarily small by increasing $M$, the left side of (9.5) $\to 0$, and the desired result follows.

(ii)⇒(iii): This is obvious.

(iii)⇒(i): By the assumed integrability of the envelope $F$, Lemma 8.16 can be employed to verify that there is a version of $\|\mathbb{P}_n - P\|_{\mathcal{F}}^*$ that converges outer almost surely to a constant. Condition (iii) implies that this constant must be zero.□

## 9.4 Donsker Preservation

In this section, we describe several techniques for building Donsker classes from other Donsker classes. The first theorem, Theorem 9.30, gives results for subsets, pointwise closures and symmetric convex hulls of Donsker classes. The second theorem, Theorem 9.31, presents a very powerful result for Lipschitz functions of Donsker classes. The corollary that follows presents consequences of this theorem that are quite useful in statistical applications.

For a class $\mathcal{F}$ of real-valued, measurable functions on the sample space $\mathcal{X}$, let $\overline{\mathcal{F}}^{(P,2)}$ be the set of all $f : \mathcal{X} \mapsto \mathbb{R}$ for which there exists a sequence $\{f_m\} \in \mathcal{F}$ such that $f_m \to f$ both pointwise (i.e., for every argument $x \in \mathcal{X}$) and in $L_2(P)$. Similarly, let $\overline{\text{sconv}}^{(P,2)}\,\mathcal{F}$ be the pointwise and $L_2(P)$ closure of $\text{sconv}\,\mathcal{F}$ defined in Section 9.1.1.

THEOREM 9.30 *Let $\mathcal{F}$ be a $P$-Donsker class. Then*

*(i) For any $\mathcal{G} \subset \mathcal{F}$, $\mathcal{G}$ is $P$-Donsker.*

*(ii) $\overline{\mathcal{F}}^{(P,2)}$ is $P$-Donsker.*

*(iii) $\overline{sconv}^{(P,2)}\,\mathcal{F}$ is $P$-Donsker.*

**Proof.** The proof of (i) is obvious by the facts that weak convergence consists of marginal convergence plus asymptotic equicontinuity and that the maximum modulus of continuity does not increase when maximizing over a smaller set. For (ii), one can without loss of generality assume that both $\mathcal{F}$ and $\overline{\mathcal{F}}^{(P,2)}$ are mean zero classes. For a class of measurable functions $\mathcal{G}$, denote the modulus of continuity

$$M_{\mathcal{G}}(\delta) \equiv \sup_{f,g \in \mathcal{G}: \|f-g\|_{P,2} < \delta} |\mathbb{G}_n(f-g)|.$$

Fix $\delta > 0$. We can choose $f, g \in \overline{\mathcal{F}}^{(P,2)}$ such that $|\mathbb{G}_n(f-g)|$ is arbitrarily close to $M_{\overline{\mathcal{F}}^{(P,2)}}(\delta)$ and $\|f-g\|_{P,2} < \delta$. We can also choose $f_*, g_* \in \mathcal{F}$ such that $\|f-f_*\|_{P,2}$ and $\|g-g_*\|_{P,2}$ are arbitrarily small (for fixed data). Thus $M_{\overline{\mathcal{F}}^{(P,2)}}(\delta) \leq M_{\mathcal{F}}(2\delta)$. Since $\delta > 0$ was arbitrary, we obtain that asymptotic equicontinuity in probability for $\overline{\mathcal{F}}^{(P,2)}$ follows from asymptotic equicontinuity in probability of $\{\mathbb{G}_n(f) : f \in \mathcal{F}\}$. Part (iii) is Theorem 2.10.3 of VW, and we omit its proof.□

The following theorem, Theorem 2.10.6 of VW, is one of the most useful Donsker preservation results for statistical applications. We omit the proof.

THEOREM 9.31 *Let $\mathcal{F}_1, \ldots, \mathcal{F}_k$ be Donsker classes with* $\max_{1 \leq i \leq k} \|P\|_{\mathcal{F}_i} < \infty$. *Let $\phi : \mathbb{R}^k \mapsto \mathbb{R}$ satisfy*

$$|\phi \circ f(x) - \phi \circ g(x)|^2 \leq c^2 \sum_{i=1}^{k} (f_i(x) - g_i(x))^2,$$

*for every $f, g \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_k$ and $x \in \mathcal{X}$ and for some constant $c < \infty$. Then $\phi \circ (\mathcal{F}_1, \ldots, \mathcal{F}_k)$ is Donsker provided $\phi \circ f$ is square integrable for at least one $f \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_k$.*

The following corollary is a useful consequence of this theorem:

COROLLARY 9.32 *Let $\mathcal{F}$ and $\mathcal{G}$ be Donsker classes. Then:*

*(i) $\mathcal{F} \cup \mathcal{G}$ and $\mathcal{F} + \mathcal{G}$ are Donsker.*

*(ii) If $\|P\|_{\mathcal{F}\cup\mathcal{G}} < \infty$, then the classes of pairwise infima, $\mathcal{F} \wedge \mathcal{G}$, and pairwise suprema, $\mathcal{F} \vee \mathcal{G}$, are both Donsker.*

*(iii) IF $\mathcal{F}$ and $\mathcal{G}$ are both uniformly bounded, $\mathcal{F} \cdot \mathcal{G}$ is Donsker.*

*(iv) If $\psi : \overline{R} \mapsto \mathbb{R}$ is Lipschitz continuous, where $R$ is the range of functions in $\mathcal{F}$, and $\|\psi(f)\|_{P,2} < \infty$ for at least one $f \in \mathcal{F}$, then $\psi(\mathcal{F})$ is Donsker.*

*(v) If $\|P\|_{\mathcal{F}} < \infty$ and $g$ is a uniformly bounded, measurable function, then $\mathcal{F} \cdot g$ is Donsker.*

**Proof.** For any measurable function $f$, let $\dot{f} \equiv f - Pf$. Also define $\dot{\mathcal{F}} \equiv \{\dot{f} : f \in \mathcal{F}\}$ and $\dot{\mathcal{G}} \equiv \{\dot{f} : f \in \mathcal{G}\}$. Note that for any $f \in \mathcal{F}$ and $g \in \mathcal{G}$, $\mathbb{G}_n f = \mathbb{G}_n \dot{f}$, $\mathbb{G}_n g = \mathbb{G}_n \dot{g}$ and $\mathbb{G}_n(f+g) = \mathbb{G}_n(\dot{f}+\dot{g})$. Hence $\mathcal{F}\cup\mathcal{G}$ is Donsker if and only if $\dot{\mathcal{F}} \cup \dot{\mathcal{G}}$ is Donsker; and, similarly, $\mathcal{F}+\mathcal{G}$ is Donsker if and only $\dot{\mathcal{F}} + \dot{\mathcal{G}}$ is Donsker. Clearly, $\|P\|_{\dot{\mathcal{F}}\cup\dot{\mathcal{G}}} = 0$. Hence Lipschitz continuity of the map $(x, y) \mapsto x + y$ on $\mathbb{R}^2$ yields that $\dot{\mathcal{F}} + \dot{\mathcal{G}}$ is Donsker, via Theorem 9.31. Hence also $\mathcal{F} + \mathcal{G}$ is Donsker. Since $\dot{\mathcal{F}} \cup \dot{\mathcal{G}}$ is contained in the union of $\dot{\mathcal{F}} \cup \{0\}$ and $\dot{\mathcal{G}} \cup \{0\}$, $\dot{\mathcal{F}} \cup \dot{\mathcal{G}}$ is Donsker and hence so is $\mathcal{F} \cup \mathcal{G}$. Thus Part (i) follows.

Proving Parts (ii) and (iv) is saved as an exercise. Part (iii) follows since the map $(x, y) \mapsto xy$ is Lipschitz continuous on bounded subsets of $\mathbb{R}^2$. For Part (v), note that for any $f_1, f_2 \in \mathcal{F}$, $|f_1(x)g(x) - f_2(x)g(x)| \leq \|g\|_\infty |f_1(x) - f_2(x)|$. Hence $\phi \circ \{\mathcal{F}, \{g\}\}$, where $\phi(x, y) = xy$, is Lipschitz continuous in the required manner.□

## 9.5 Proofs

**Proof of Theorem 9.3.** Let $\mathcal{C}$ denote the set of all subgraphs $C_f$ of functions $f \in \mathcal{F}$. Note that for any probability measure $Q$ on $\mathcal{X}$ and any $f, g \in \mathcal{F}$,

$$\begin{aligned} Q|f - g| &= \int_{\mathcal{X}} \int_{\mathbb{R}} |1\{t < f(x)\} - 1\{t < g(x)\}| \, dt \, Q(dx) \\ &= (Q \times \lambda)(C_f \Delta C_g), \end{aligned}$$

where $\lambda$ is Lebesgue measure, $A\Delta B \equiv A\cup B - A\cap B$ for any two sets $A, B$, and the second equality follows from Fubini's theorem. Construct a probability measure $P$ on $\mathcal{X}\times\mathbb{R}$ by restricting $Q\times\lambda$ to the set $\{(x,t) : |t| \le F(x)\}$ and letting $P = (Q\times\lambda)/(2\|F\|_{Q,1})$. Now $Q|f-g| = 2\|F\|_{Q,1}P|1\{C_f\} - 1\{C_g\}|$. Thus, by Theorem 9.2 above,

$$\begin{aligned} N(\epsilon 2\|F\|_{Q,1}, \mathcal{F}, L_1(Q)) &= N(\epsilon, \mathcal{C}, L_1(P)) \\ &\le KV(\mathcal{C})(4e)^{V(\mathcal{C})}\left(\frac{1}{\epsilon}\right)^{V(\mathcal{C})-1}. \end{aligned} \tag{9.6}$$

For $r > 1$, we have

$$Q|f-g|^r \le Q\left\{|f-g|(2F)^{r-1}\right\} = 2^{r-1}R|f-g|QF^{r-1},$$

where $R$ is the probability measure with density $F^{r-1}/QF^{r-1}$ with respect to $Q$. Thus

$$\begin{aligned} \|f-g\|_{Q,r} &\le 2^{1-1/r}\|f-g\|_{R,1}^{1/r}\left(QF^{r-1}\right)^{1/r} \\ &= 2^{1-1/r}\|f-g\|_{R,1}^{1/r}\|F\|_{Q,r}\left(\frac{QF^{r-1}}{QF^r}\right)^{1/r}, \end{aligned}$$

which implies

$$\frac{\|f-g\|_{Q,r}}{2\|F\|_{Q,r}} \le \left(\frac{\|f-g\|_{R,1}}{2\|F\|_{R,1}}\right)^{1/r}.$$

Hence $N(\epsilon 2\|F\|_{Q,r}, \mathcal{F}, L_r(Q)) \le N(\epsilon^r 2\|F\|_{R,1}, \mathcal{F}, L_1(R))$. Since (9.6) applies equally well with $Q$ replaced by $R$, we now have that

$$N(\epsilon 2\|F\|_{Q,r}, \mathcal{F}, L_r(Q)) \le KV(\mathcal{C})(4e)^{V(\mathcal{C})}\left(\frac{1}{\epsilon}\right)^{r(V(\mathcal{C})-1)}.$$

The desired result now follows by bringing the factor 2 in the left-hand-side over to the numerator of $1/\epsilon$ in the right-hand-side.□

**Proof of Lemma 9.9.** We leave it as an exercise to show that Parts (i) and (ii) follow from Parts (ii) and (iii) of Lemma 9.7. To prove (iii), note first that the sets $\{x : f(x) > 0\}$ are (obviously) one-to-one images of the sets $\{(x,0) : f(x) > 0\}$, which are the intersections of the open subgraphs $\{(x,t) : f(x) > t\}$ with the set $\mathcal{X}\times\{0\}$. Since a single set has VC index 1, the result now follows by applying (ii) and then (iv) of Lemma 9.7. The subgraphs of $-\mathcal{F}$ are the images of the *open supergraphs* $\{(x,t) : t > f(x)\}$ under the one-to-one map $(x,t)\mapsto(x,-t)$. Since the open supergraphs are the complements of the *closed subgraphs* $\{(x,t) : t \ge f(x)\}$, they have the same VC-index as $\mathcal{F}$ by Lemma 9.33 below and by Part (i) of Lemma 9.7.

Part (v) follows from the fact that $\mathcal{F}+g$ shatters a given set of points $(x_1,t_1),\ldots,(x_n,t_n)$ if and only if $\mathcal{F}$ shatters $(x_1,t_1'),\ldots,(x_n,t_n')$, where

$t_i' = t_i - g(x_i)$, $1 \le i \le n$. For Part (vi), note that for any $f \in \mathcal{F}$ the subgraph of $fg$ is the union of the sets $C_f^+ \equiv \{(x,t) : t < f(x)g(x), g(x) > 0\}$, $C_f^- \equiv \{(x,t) : t < f(x)g(x), g(x) < 0\}$ and $C_f^0 \equiv \{(x,t) : t < 0, g(x) = 0\}$. Define $\mathcal{C}^+ \equiv \{C_f^+ : f \in \mathcal{F}\}$, $\mathcal{C}^- \equiv \{C_f^- : f \in \mathcal{F}\}$ and $\mathcal{C}^0 \equiv \{C_f^0 : f \in \mathcal{F}\}$. By Exercise 9.6.6 below, it suffices to show that these three classes are VC on the respective disjoint sets $\mathcal{X} \cap \{x : g(x) > 0\} \times \mathbb{R}$, $\mathcal{X} \cap \{x : g(x) < 0\} \times \mathbb{R}$ and $\mathcal{X} \cap \{x : g(x) = 0\} \times \mathbb{R}$, with respective VC indices bounded by $V_{\mathcal{F}}$, $V_{\mathcal{F}}$ and 1. Consider first $\mathcal{C}^+$ on $\mathcal{X} \cap \{x : g(x) < 0\}$. Note that the subset $(x_1, t_1), \ldots, (x_m, t_m)$ is shattered by $\mathcal{C}^+$ if and only if the subset $(x_1, t_1/g(x_1)), \ldots, (x_m, t_m/g(x_m))$ is shattered by the subgraphs of $\mathcal{F}$. Thus the VC-index of $\mathcal{C}^+$ on the relevant subset of $\mathcal{X} \times \mathbb{R}$ is $V_{\mathcal{F}}$. The same VC-index occurs for $\mathcal{C}^-$, but the VC-index for $\mathcal{C}^0$ is clearly 1. This concludes the proof of (vi).

For (vii), the result follows from Part (vi) of Lemma 9.7 since the subgraphs of the class $\mathcal{F} \circ \psi$ are the inverse images of the subgraphs of $\mathcal{F}$ under the map $(z,t) \mapsto (\psi(z), t)$. For Part (viii), suppose that the subgraphs of $\phi \circ \mathcal{F}$ shatter the set of points $(x_1, t_1), \ldots, (x_n, t_n)$. Now choose $f_1, \ldots, f_m$ from $\mathcal{F}$ so that the subgraphs of the functions $\phi \circ f_j$ pick out all $m = 2^n$ subsets. For each $1 \le i \le n$, define $s_i$ to be the largest value of $f_j(x_i)$ over those $j \in \{1, \ldots, m\}$ for which $\phi(f_j(x_i)) \le t_i$. Now note that $f_j(x_i) \le s_i$ if and only if $\phi(f_j(x_i)) \le t_i$, for all $1 \le i \le n$ and $1 \le j \le m$. Hence the subgraphs of $f_1, \ldots, f_m$ shatter the points $(x_1, s_1), \ldots, (x_n, s_n)$.□

**Proof of Lemma 9.11.** First consider the class $\mathcal{H}_{+,r}$ of monotone increasing, right-continuous functions $h : \mathbb{R} \mapsto [0,1]$. For each $h \in \mathcal{H}_{+,r}$, define $h^{-1}(t) \equiv \inf\{x : h(x) \ge t\}$, and note that for any $x, t \in \mathbb{R}$, $h(x) \ge t$ if and only if $x \ge h^{-1}(t)$. Thus the class of indicator functions $\{1\{h(x) \ge t\} : h \in \mathcal{H}_{+,r}\} = \left\{1\{x \ge h^{-1}(t)\} : h \in \mathcal{H}_{+,r}\right\} \subset \{1\{x \ge t\} : t \in \mathbb{R}\}$. Since the last class of sets has VC index 2, the first class is also VC with index 2. Since each function $h \in \mathcal{H}_{+,r}$ is the pointwise limit of the sequence

$$h_m = \sum_{j=1}^{m} \frac{1}{m} 1\left\{h(x) \ge \frac{j}{m}\right\},$$

we have that $\mathcal{H}_{+,r}$ is contained in the closed convex hull of a VC-subgraph class with VC index 2. Thus, by Corollary 9.5, we have for all $0 < \epsilon < 1$,

$$\sup_Q \log N(\epsilon, \mathcal{H}_{+,r}, L_2(Q)) \le \frac{K_0}{\epsilon},$$

where the supremum is taken over all probability measures $Q$ and the constant $K_0$ is universal. Now consider the class $\mathcal{H}_{+,l}$ of monotone increasing, left-continuous functions, and define $\tilde{h}^{-1}(x) \equiv \sup\{x : h(x) \le t\}$. Now note that for any $x, t \in \mathbb{R}$, $h(x) > t$ if and only if $x > \tilde{h}^{-1}(t)$. Arguing as before, we deduce that $\{1\{h(x) > t\} : h \in \mathcal{H}_{+,l}\}$ is a VC-class with index 2. Since each $h \in \mathcal{H}_{+,l}$ is the pointwise limit of the sequence

$$h_m = \sum_{j=1}^{m} \frac{1}{m} 1\left\{h(x) > \frac{j}{m}\right\},$$

we can again apply Corollary 9.5 to arrive at the same uniform entropy bound we arrived at for $\mathcal{H}_{+,r}$.

Now let $\mathcal{H}_+$ be the class of all monotone increasing functions $h : \mathbb{R} \mapsto [0,1]$, and note that each $h \in \mathcal{H}_+$ can be written as $h_r + h_l$, where $h_r \in \mathcal{H}_{+,r}$ and $h_l \in \mathcal{H}_{+,l}$. Hence for any probability measure $Q$ and any $h^{(1)}, h^{(2)} \in \mathcal{H}_+$, $\|h^{(1)} - h^{(2)}\|_{Q,2} \leq \|h_r^{(1)} - h_r^{(2)}\|_{Q,r} + \|h_l^{(1)} - h_l^{(2)}\|_{Q,2}$, where $h_r^{(1)}, h_r^{(2)} \in \mathcal{H}_{+,r}$ and $h_l^{(1)}, h_l^{(2)} \in \mathcal{H}_{+,l}$. Thus $N(\epsilon, \mathcal{H}_+, L_2(Q)) \leq N(\epsilon/2, \mathcal{H}_{+,r}, L_2(Q))$ $\times N(\epsilon/2, \mathcal{H}_{+,l}, L_2(Q))$, and hence

$$\sup_Q \log N(\epsilon, \mathcal{H}_+, L_2(Q)) \leq \frac{K_1}{\epsilon},$$

where $K_1 = 4K_0$. Since any monotone decreasing function $g : \mathbb{R} \mapsto [0,1]$ can be written as $1 - h$, where $h \in \mathcal{H}_+$, the uniform entropy numbers for the class of all monotone functions $f : \mathbb{R} \mapsto [0,1]$, which we denote $\mathcal{F}$, is $\log(2)$ plus the uniform entropy numbers for $\mathcal{H}_+$. Since $0 < \epsilon < 1$, we obtain the desired conclusion given in the statement of the lemma, with $K = \sqrt{2}K_1 = \sqrt{32}K_0$.□

LEMMA 9.33 *Let $\mathcal{F}$ be a set of measurable functions on $\mathcal{X}$. Then the closed subgraphs have the same VC-index as the open subgraphs.*

**Proof.** Suppose the closed subgraphs (the subgraphs of the form $\{(x,t) : t \leq f(x)\}$) shatter the set of points $(x_1,t_1),\ldots,(x_n,t_n)$. Now select out of $\mathcal{F}$ functions $f_1,\ldots,f_m$ whose closed subgraphs shatter all $m = 2^n$ subsets. Let $\epsilon = (1/2)\inf\{t_i - f_j(x_i) : t_i - f_j(x_i) > 0\}$, and note that the open subgraphs (the subgraphs of the form $\{(x,t), t < f(x)\}$) of the $f_1,\ldots,f_m$ shatter the set of points $(x_1,t_1-\epsilon),\ldots,(x_n,t_n-\epsilon)$. This follows since, by construction, $t_i - \epsilon \geq f_j(x_i)$ if and only if $t_i > f_j(x_i)$, for all $1 \leq i \leq n$ and $1 \leq j \leq m$. Now suppose the open subgraphs shatter the set of points $(x_1,t_1),\ldots,(x_n,t_n)$. Select out of $\mathcal{F}$ functions $f_1,\ldots,f_m$ whose open subgraphs shatter all $m = 2^n$ subsets, and let $\epsilon = (1/2)\inf\{f_j(x_i) - t_i : f_j(x_i) - t_i > 0\}$. Note now that the closed subgraphs of $f_1,\ldots,f_m$ shatter the set of points $(x_1,t_1+\epsilon),\ldots,(x_n,t_n+\epsilon)$, since, by construction, $t_i < f_j(x_i)$ if and only if $t_i + \epsilon \leq f_j(x_i)$. Thus the VC-indices of open and closed subgraphs are the same.□

## 9.6 Exercises

**9.6.1.** Show that $\text{sconv}\mathcal{F} \subset \text{conv}\,\mathcal{G}$, where $\mathcal{G} = \mathcal{F} \cup \{-\mathcal{F}\} \cup \{0\}$.

**9.6.2.** Show that the expression $N(\epsilon\|aF\|_{bQ,r}, a\mathcal{F}, L_r(bQ))$ does not depend on the constants $0 < a, b < \infty$, where $1 \leq r < \infty$.

**9.6.3.** Prove Corollary 9.5.

**9.6.4.** In the proof of Lemma 9.7, verify that Part (iii) follows from Parts (i) and (ii).

**9.6.5.** Show that Parts (i) and (ii) of Lemma 9.9 follow from Parts (ii) and (iii) of Lemma 9.7.

**9.6.6.** Let $\mathcal{X} = \cup_{i=1}^m \mathcal{X}_i$, where the $\mathcal{X}_i$ are disjoint; and assume $\mathcal{C}_i$ is a VC-class of subsets of $\mathcal{X}_i$, with VC-index $V_i$, $1 \leq i \leq m$. Show that $\sqcup_{i=1}^m \mathcal{C}_i$ is a VC-class in $\mathcal{X}$ with VC-index $V_1 + \cdots + V_m - m + 1$. Hint: Note that $\mathcal{C}_1 \cup \mathcal{X}_2$ and $\mathcal{X}_1 \cup \mathcal{C}_2$ are VC on $\mathcal{X}_1 \cup \mathcal{X}_2$ with respective indices $V_1$ and $V_2$. Now use Part (ii)—not Part (iii)—of Lemma 9.7 to show that $\mathcal{C}_1 \sqcup \mathcal{C}_2$ is VC on $\mathcal{X}_1 \cup \mathcal{X}_2$ with VC index $V_1 + V_2 - 1$.

**9.6.7.** Show that if $\mathcal{F}$ and $\mathcal{G}$ are BUEI with respective envelopes $F$ and $G$, then $\mathcal{F} \sqcup \mathcal{G}$ is BUEI with envelope $F \vee G$.

**9.6.8.** In the context of the simple linear regression example of Section 4.4.1, verify the following:

(a) Show that both $\mathcal{G}_1$ and $\mathcal{G}_2$ are Donsker even though neither $U$ nor $e$ are bounded. Hint: Use BUEI preservation results.

(b) Verify that both

$$\sup_{z \in [a+h, b-h]} \left| \int_{\mathbb{R}} h^{-1} L\left(\frac{z-u}{h}\right) H(du) - \dot{H}(z) \right| = O(h)$$

and

$$\left( \sup_{z \in [a, a+h)} \left| \dot{H}(z) - \dot{H}(a+h) \right| \right) \vee \left( \sup_{z \in (b-h, b]} \left| \dot{H}(z) - \dot{H}(b-h) \right| \right)$$

$$= O(h).$$

(c) Show that $\mathcal{F}_1$ is Donsker and $\mathcal{F}_2$ is Glivenko-Cantelli.

**9.6.9.** Prove Lemma 9.25.

**9.6.10.** Consider the class $\mathcal{F}$ of all functions $f : [0,1] \mapsto [0,1]$ such that $|f(x) - f(y)| \leq |x - y|$. Show that a set of $\epsilon$-brackets can be constructed to cover $\mathcal{F}$ with cardinality bounded by $\exp(C/\epsilon)$ for some $0 < C < \infty$. Hint: Fix $\epsilon > 0$, and let $n$ be the smallest integer $\geq 3/\epsilon$. For any $p = (k_0, \ldots, k_n) \in P \equiv \{1, \ldots, n\}^{n+1}$, define the path $\bar{p}$ to be the collection of all function in $\mathcal{F}$ such that $f \in \bar{p}$ only if $f(i/n) \in [(k_i - 1)/n, k_i/n]$ for all $i = 0 \ldots n$. Show that for all $f \in \mathcal{F}$, if $f(i/n) \in [j/n, (j+1)/n]$, then

$$f\left[\frac{i+1}{n}\right] \in \left[\frac{(j-1)\vee 0}{n}, \frac{(j+2)\wedge n}{n}\right]$$

for $i, j = 0, \ldots, n-1$. Show that this implies that the number of paths of the form $\bar{p}$ for $p \in P$ needed to "capture" all elements of $\mathcal{F}$ is bounded by $n3^n$. Now show that for each $p \in P$, there exists a pair of right-continuous "bracketing" functions $L_p, U_p : [0,1] \mapsto [0,1]$ such that $\forall x \in [0,1]$, $L_p(x) < U_p(x)$, $U_p(x) - L_p(x) \leq 3/n \leq \epsilon$, and $L_p(x) \leq f(x) \leq U_p(x)$ for all $f \in \bar{p}$. Now complete the proof.

**9.6.11.** Show that if $\mathcal{F}$ is Donsker with $\|P\|_{\mathcal{F}} < \infty$ and $f \geq \delta$ for all $f \in \mathcal{F}$ and some constant $\delta > 0$, then $1/\mathcal{F} \equiv \{1/f : f \in \mathcal{F}\}$ is Donsker.

**9.6.12.** Complete the proof of Corollary 9.32:

1. Prove Part (ii). Hint: show first that for any real numbers $a_1, a_2, b_1, b_2$, $|a_1 \wedge b_1 - a_2 \wedge b_2| \leq |a_1 - a_2| + |b_1 - b_2|$.
2. Prove Part (iv).

## 9.7 Notes

Theorem 9.3 is a minor modification of Theorem 2.6.7 of VW. Corollary 9.5, Lemma 9.6 and Theorem 9.23 are Corollary 2.6.12, Lemma 2.6.15 and Theorem 2.7.11, respectively, of VW. Lemmas 9.7 and 9.9 are modification of Lemmas 2.6.17 and 2.6.18, respectively, of VW. Lemma 9.11 was suggested by Example 2.6.21 of VW, and Lemma 9.12 is a modification of Exercise 2.6.14 of VW. Parts (i) and (ii) of Theorem 9.30 are Theorems 2.10.1 and 2.10.2, respectively, of VW. Corollary 9.32 includes some modifications of Examples 2.10.7, 2.10.8 and 2.10.10 of VW. Lemma 9.33 was suggested by Exercise 2.6.10 of VW. Exercise 9.6.10 is a modification of Exercise 19.5 of van der Vaart (1998).

The bounded uniform entropy integral (BUEI) preservation techniques presented here grew out of the author's work on estimating equations for functional data described in Fine, Yan and Kosorok (2004).

# 10
# Bootstrapping Empirical Processes

The purpose of this chapter is to obtain consistency results for bootstrapped empirical processes. These results can then be applied to many kinds of bootstrapped estimators since most estimators can be expressed as functionals of empirical processes. Much of the bootstrap results for such estimators will be deferred to later chapters where we discuss the functional delta method, Z-estimation and M-estimation. We do, however, present one specialized result for parametric Z-estimators in Section 3 of this chapter as a practical illustration of bootstrap techniques.

We note that both conditional and unconditional bootstrap consistency results can be useful depending on the application. For the conditional bootstrap, the goal is to establish convergence of the conditional law given the data to an unconditional limit law. This convergence can be either in probability or outer almost sure. While the later convergence is certainly stronger, convergence in probability is usually sufficient for statistical applications.

The best choice of bootstrap weights for a given statistical application is also an important question, and the answer depends on the application. While the multinomial bootstrap is conceptually simple, its use in survival analysis applications may result in too much tied data. In the presence of censoring, it is even possible that a bootstrap sample could be drawn that consists of only censored observations. To avoid complications of this kind, it may be better to use the Bayesian bootstrap (Rubin, 1981). The weights for the Bayesian bootstrap are $\xi_1/\bar{\xi}, \ldots, \xi_n/\bar{\xi}$, where $\xi_1, \ldots, \xi_n$ are i.i.d. standard exponential (mean and variance 1), independent of the data $X_1, \ldots, X_n$, and where $\bar{\xi} \equiv n^{-1} \sum_{i=1}^{n} \xi_i$. Since these weights are strictly

positive, all observations are represented in each bootstrap realization, and the aforementioned problem with tied data won't happen unless the original data has ties. Both the multinomial and Bayesian bootstraps are included in the bootstrap weights we discuss in this chapter.

The multinomial weighted bootstrap is sometimes called the *nonparametric bootstrap* since it amounts to sampling from the empirical distribution, which is a nonparametric estimate of the true distribution. In contrast, the *parametric bootstrap* is obtained by sampling from a parametric estimate $P_{\hat{\theta}_n}$ of the true distribution, where $\hat{\theta}_n$ is a consistent estimate of the true value of $\theta$ (see, for example, Chapter 1 of Shao and Tu, 1995). A detailed discussion of the parametric bootstrap is beyond the scope of this chapter. Another kind of bootstrap is the exchangeable weighted bootstrap, which we only mention briefly in Lemma 10.18 below. This lemma is needed for the proof of Theorem 10.15.

We also note that the asymptotic results of this chapter are all first order, and in this situation the limiting results do not vary among those schemes that satisfy the stated conditions. A more refined analysis of differences between weighting schemes is beyond the scope of this chapter, but such differences may be important in small samples. A good reference for higher order properties of the bootstrap is Hall (1992).

The first section of this chapter considers unconditional and conditional convergence of bootstrapped empirical processes to limiting laws when the class of functions involved is Donsker. The main result of this section is a proof of Theorems 2.6 and 2.7 on Page 20 of Section 2.2.3. At the end of the section, we present several special continuous mapping results for bootstrapped processes. The second section considers parallel results when the function class involved is Glivenko-Cantelli. In this case, the limiting laws are degenerate, i.e., constant with probability 1. Such results are helpful for establishing consistency of bootstrapped estimators. The third section presents the simple Z-estimator illustration promised above. Throughout this chapter, we will sometimes for simplicity omit the subscript when referring to a representative of an i.i.d. sample. For example, we may use $\mathrm{E}|\xi|$ to refer to $\mathrm{E}|\xi_1|$, where $\xi_1$ is the first member of the sample $\xi_1, \ldots, \xi_n$. The context will make the meaning clear.

## 10.1 The Bootstrap for Donsker Classes

The overall goal of this section is to prove the validity of the bootstrap central limit theorems given in Theorems 2.6 and 2.7 on Page 20 of Chapter 2. Both unconditional and conditional multiplier central limit theorems play a pivotal role in this development and will be presented first. At the end of the section, we also present several special continuous mapping results which apply to bootstrapped processes. These results allow the construction of

asymptotically uniformly valid confidence bands for $\{Pf : f \in \mathcal{F}\}$ when $\mathcal{F}$ is Donsker.

### *10.1.1 An Unconditional Multiplier Central Limit Theorem*

In this section, we present a multiplier central limit theorem that forms the basis for proving the unconditional central limit theorems of the next section. We also present an interesting corollary. For a real random variable $\xi$, recall from Section 2.2.3 the quantity $\|\xi\|_{2,1} \equiv \int_0^\infty \sqrt{\mathrm{P}(|\xi| > x)}dx$. Exercise 10.5.1 below verifies this is a norm which is slightly larger than $\|\cdot\|_2$. Also recall that $\delta_{X_i}$ is the probability measure that assigns a mass of 1 to $X_i$ so that $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ and $\mathbb{G}_n = n^{-1/2}\sum_{i=1}^n(\delta_{X_i} - P)$.

THEOREM 10.1 *(Multiplier central limit theorem) Let $\mathcal{F}$ be a class of measurable functions, and let $\xi_1, \ldots, \xi_n$ be i.i.d. random variables with mean zero, variance 1, and with $\|\xi\|_{2,1} < \infty$, independent of the sample data $X_1, \ldots, X_n$. Let $\mathbb{G}'_n \equiv n^{-1/2}\sum_{i=1}^n \xi_i(\delta_{X_i} - P)$ and $\mathbb{G}''_n \equiv n^{-1/2}\sum_{i=1}^n (\xi_i - \bar{\xi})\delta_{X_i}$, where $\bar{\xi} \equiv n^{-1}\sum_{i=1}^n \xi_i$. Then the following are equivalent:*

*(i) $\mathcal{F}$ is P-Donsker;*

*(ii) $\mathbb{G}'_n$ converges weakly to a tight process in $\ell^\infty(\mathcal{F})$;*

*(iii) $\mathbb{G}'_n \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathcal{F})$;*

*(iv) $\mathbb{G}''_n \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathcal{F})$.*

Before giving the proof of this theorem, we will need the following tool. This lemma is Lemma 2.9.1 of VW, and we give it without proof:

LEMMA 10.2 *(Multiplier inequalities) Let $Z_1, \ldots, Z_n$ be i.i.d. stochastic processes, with index $\mathcal{F}$ such that $\mathrm{E}^*\|Z\|_\mathcal{F} < \infty$, independent of the i.i.d. Rademacher variables $\epsilon_1, \ldots, \epsilon_n$. Then for every i.i.d. sample $\xi_1, \ldots, \xi_n$ of real, mean-zero random variables independent of $Z_1, \ldots, Z_n$, and any $1 \le n_0 \le n$,*

$$\begin{aligned}
\frac{1}{2}\|\xi\|_1 \mathrm{E}^* \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i Z_i \right\|_\mathcal{F} &\le \mathrm{E}^* \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i Z_i \right\|_\mathcal{F} \\
&\le 2(n_0 - 1)\mathrm{E}^*\|Z\|_\mathcal{F} \mathrm{E} \max_{1 \le i \le n} \frac{|\xi_i|}{\sqrt{n}} \\
&\quad + 2\sqrt{2}\|\xi\|_{2,1} \max_{n_0 \le k \le n} \mathrm{E}^* \left\| \frac{1}{\sqrt{k}} \sum_{i=n_0}^k \epsilon_i Z_i \right\|_\mathcal{F}.
\end{aligned}$$

*When the $\xi_i$ are symmetrically distributed, the constants $1/2$, 2 and $2\sqrt{2}$ can all be replaced by 1.*

**Proof of Theorem 10.1.** Note that the processes $\mathbb{G}$, $\mathbb{G}_n$, $\mathbb{G}'_n$ and $\mathbb{G}''_n$ do not change if they are indexed by $\dot{\mathcal{F}} \equiv \{f - Pf : f \in \mathcal{F}\}$ rather than $\mathcal{F}$. Thus we can assume throughout the proof that $\|P\|_{\mathcal{F}} = 0$ without loss of generality.

(i)$\Leftrightarrow$(ii): Convergence of the finite-dimensional marginal distributions of $\mathbb{G}_n$ and $\mathbb{G}'_n$ is equivalent to $\mathcal{F} \subset L_2(P)$, and thus it suffices to show that the asymptotic equicontinuity conditions of both processes are equivalent. By Lemma 8.17, if $\mathcal{F}$ is Donsker, then $P^*(F > x) = o(x^{-2})$ as $x \to \infty$. Similarly, if $\xi \cdot \mathcal{F}$ is Donsker, then $P^*(|\xi| \times F > x) = o(x^{-2})$ as $x \to \infty$. In both cases, $P^*F < \infty$. Since the variance of $\xi$ is finite, we have by Exercise 10.5.2 below that $\mathrm{E}^* \max_{1 \le i \le n} |\xi_i|/\sqrt{n} \to 0$. Combining this with the multiplier inequality (Lemma 10.2), we have

$$\frac{1}{2}\|\xi\|_1 \limsup_{n\to\infty} \mathrm{E}^* \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f(X_i) \right\|_{\mathcal{F}_\delta} \le \limsup_{n\to\infty} \mathrm{E}^* \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i f(X_i) \right\|_{\mathcal{F}_\delta}$$
$$\le 2\sqrt{2}\|\xi\|_{2,1} \sup_{k \ge n_0} \mathrm{E}^* \left\| \frac{1}{\sqrt{k}} \sum_{i=1}^k \epsilon_i f(X_i) \right\|_{\mathcal{F}_\delta},$$

for every $\delta > 0$ and $n_0 \le n$. By the symmetrization theorem (Theorem 8.8), we can remove the Rademacher variables $\epsilon_1, \ldots, \epsilon_n$ at the cost of changing the constants. Hence, for any sequence $\delta_n \downarrow 0$, $\mathrm{E}^* \|n^{-1/2} \sum_{i=1}^n (\delta_{X_i} - P)\|_{\mathcal{F}_{\delta_n}} \to 0$ if and only if $\mathrm{E}^* \|n^{-1/2} \sum_{i=1}^n \xi_i (\delta_{X_i} - P)\|_{\mathcal{F}_{\delta_n}} \to 0$. By Lemma 8.17, these mean versions of the asymptotic equicontinuity conditions imply the probability versions, and the desired results follow. We have actually proved that the first three assertions are equivalent.

(iii)$\Rightarrow$(iv): Note that by the equivalence of (i) and (iii), $\mathcal{F}$ is Glivenko-Cantelli. Since $\mathbb{G}'_n - \mathbb{G}''_n = \sqrt{n}\bar{\xi}\mathbb{P}_n$, we now have that $\|\mathbb{G}'_n - \mathbb{G}''_n\|_{\mathcal{F}} \xrightarrow{\mathrm{P}} 0$. Thus (iv) follows.

(iv)$\Rightarrow$(i): Let $(Y_1, \ldots, Y_n)$ be an independent copy of $(X_1, \ldots, X_n)$, and let $(\tilde{\xi}_1, \ldots, \tilde{\xi}_n)$ be an independent copy of $(\xi_1, \ldots, \xi_n)$, so that $(\xi_1, \ldots, \xi_n, \tilde{\xi}_1, \ldots, \tilde{\xi}_n)$ is independent of $(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$. Let $\bar{\xi}$ be the pooled mean of the $\xi_i$s and $\tilde{\xi}_i$s; set

$$\mathbb{G}''_{2n} = (2n)^{-1/2} \left( \sum_{i=1}^n (\xi_i - \bar{\xi})\delta_{X_i} + \sum_{i=1}^n (\tilde{\xi}_i - \bar{\xi})\delta_{Y_i} \right)$$

and define

$$\tilde{\mathbb{G}}''_{2n} \equiv (2n)^{-1/2} \left( \sum_{i=1}^n (\tilde{\xi}_i - \bar{\xi})\delta_{X_i} + \sum_{i=1}^n (\xi_i - \bar{\xi})\delta_{Y_i} \right).$$

We now have that both $\mathbb{G}''_{2n} \rightsquigarrow \mathbb{G}$ and $\tilde{\mathbb{G}}''_{2n} \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathcal{F})$.

Thus, by the definition of weak convergence, we have that $(\mathcal{F}, \rho_P)$ is totally bounded and that for any sequence $\delta_n \downarrow 0$ both $\|\mathbb{G}''_{2n}\|_{\mathcal{F}_{\delta_n}} \xrightarrow{\mathrm{P}} 0$ and

$\|\tilde{\mathbb{G}}''_{2n}\|_{\mathcal{F}_{\delta_n}} \xrightarrow{P} 0$. Hence also $\|\mathbb{G}''_{2n} - \tilde{\mathbb{G}}''_{2n}\|_{\mathcal{F}_{\delta_n}} \xrightarrow{P} 0$. However, since

$$\mathbb{G}''_{2n} - \tilde{\mathbb{G}}''_{2n} = n^{-1/2} \sum_{i=1}^{n} \frac{(\xi_i - \tilde{\xi}_i)}{\sqrt{2}} (\delta_{X_i} - \delta_{Y_i}),$$

and since the weights $\check{\xi}_i \equiv (\xi_i - \tilde{\xi}_i)/\sqrt{2}$ satisfy the moment conditions for the theorem we are proving, we now have the $\check{\mathbb{G}}_n \equiv n^{-1/2} \sum_{i=1}^{n} (\delta_{X_i} - \delta_{Y_i}) \rightsquigarrow \sqrt{2}\mathbb{G}$ in $\ell^\infty(\mathcal{F})$ by the already proved equivalence between (iii) and (i). Thus, for any sequence $\delta_n \downarrow 0$, $\mathrm{E}^* \|\check{\mathbb{G}}_n\|_{\mathcal{F}_{\delta_n}} \to 0$. Since also

$$\mathrm{E}_Y \left| \sum_{i=1}^{n} f(X_i) - f(Y_i) \right| \geq \left| \sum_{i=1}^{n} f(X_i) - \mathrm{E} f(Y_i) \right| = \left| \sum_{i=1}^{n} f(X_i) \right|,$$

we can invoke Fubini's theorem (Lemma 6.14) to yield

$$\mathrm{E}^* \|\check{\mathbb{G}}_n\|_{\mathcal{F}_{\delta_n}} \geq \mathrm{E}^* \|\mathbb{G}_n\|_{\mathcal{F}_{\delta_n}} \to 0.$$

Hence $\mathcal{F}$ is Donsker.□

We now present the following interesting corollary which shows the possibly unexpected result that the multiplier empirical process is asymptotically independent of the usual empirical process, even though the same data $X_1, \ldots, X_n$ are used in both processes:

COROLLARY 10.3 *Assume the conditions of Theorem 10.1 hold and that $\mathcal{F}$ is Donsker. Then $(\mathbb{G}_n, \mathbb{G}'_n, \mathbb{G}''_n) \rightsquigarrow (\mathbb{G}, \mathbb{G}', \mathbb{G}',)$ in $[\ell^\infty(\mathcal{F})]^3$, where $\mathbb{G}$ and $\mathbb{G}'$ are independent $P$-Brownian bridges.*

**Proof.** By the preceding theorem, the three processes are asymptotically tight marginally and hence asymptotically tight jointly. Since the first process is uncorrelated with the second process, the limiting distribution of the first process is independent of the limiting distribution of the second process. As argued in the proof of the multiplier central limit theorem, the uniform difference between $\mathbb{G}'_n$ and $\mathbb{G}''_n$ goes to zero in probability, and thus the remainder of the corollary follows.□

### *10.1.2 Conditional Multiplier Central Limit Theorems*

In this section, the convergence properties of the multiplier processes in the previous section are studied conditional on the data. This yields in-probability and outer-almost-sure conditional multiplier central limit theorems. These results are one step closer to the bootstrap validity results of the next section. For a metric space $(\mathbb{D}, d)$, define $BL_1(\mathbb{D})$ to be the space of all functions $f : \mathbb{D} \mapsto \mathbb{R}$ with Lipschitz norm bounded by 1, i.e., $\|f\|_\infty \leq 1$ and $|f(x) - f(y)| \leq d(x, y)$ for all $x, y \in \mathbb{D}$. In the current set-up, $\mathbb{D} = \ell^\infty(\mathcal{F})$, for some class of measurable functions $\mathcal{F}$, and $d$ is the corresponding uniform metric. As we did in Section 2.2.3, we will use $BL_1$ as

shorthand for $BL_1(\ell^\infty(\mathcal{F}))$. The conditional weak convergence arrows we use in Theorems 10.4 and 10.6 below were also defined in Section 2.2.3.

We now present the in-probability conditional multiplier central limit theorem:

THEOREM 10.4 *Let $\mathcal{F}$ be a class of measurable functions, and let $\xi_1, \ldots, \xi_n$ be i.i.d. random variables with mean zero, variance 1, and $\|\xi\|_{2,1} < \infty$, independent of the sample data $X_1, \ldots, X_n$. Let $\mathbb{G}'_n$, $\mathbb{G}''_n$ and $\bar{\xi}$ be as defined in Theorem 10.1. Then the following are equivalent:*

*(i) $\mathcal{F}$ is Donsker;*

*(ii) $\mathbb{G}'_n \overset{\mathrm{P}}{\underset{\xi}{\rightsquigarrow}} \mathbb{G}$ in $\ell^\infty(\mathcal{F})$ and $\mathbb{G}'_n$ is asymptotically measurable.*

*(iii) $\mathbb{G}''_n \overset{\mathrm{P}}{\underset{\xi}{\rightsquigarrow}} \mathbb{G}$ in $\ell^\infty(\mathcal{F})$ and $\mathbb{G}''_n$ is asymptotically measurable.*

Before giving the proof of this theorem, we make a few points and present Lemma 10.5 below to aid in the proof. In the above theorem, $\mathrm{E}_\xi$ denotes taking the expectation conditional on $X_1, \ldots, X_n$. Note that for a continuous function $h : \ell^\infty(\mathcal{F}) \mapsto \mathbb{R}$, if we fix $X_1, \ldots, X_n$, then $(a_1, \ldots, a_n) \mapsto h(n^{-1/2} \sum_{i=1}^n a_i(\delta_{X_i} - P))$ is a measurable map from $\mathbb{R}^n$ to $\mathbb{R}$, provided $\|f(X) - Pf\|^*_{\mathcal{F}} < \infty$ almost surely. This last inequality is tacitly assumed so that the empirical processes under investigation reside in $\ell^\infty(\mathcal{F})$. Thus the expectation $\mathrm{E}_\xi$ in conclusions (ii) and (iii) is proper. The following lemma is a conditional multiplier central limit theorem for i.i.d. Euclidean data:

LEMMA 10.5 *Let $Z_1, \ldots, Z_n$ be i.i.d. Euclidean random vectors, with $\mathrm{E}Z = 0$ and $\mathrm{E}\|Z\|^2 < \infty$, independent of the i.i.d. sequence of real random variables $\xi_1, \ldots, \xi_n$ with $\mathrm{E}\xi = 0$ and $\mathrm{E}\xi^2 = 1$. Then, conditionally on $Z_1, Z_2, \ldots$, $n^{-1/2} \sum_{i=1}^n \xi_i Z_i \rightsquigarrow N(0, \mathrm{cov}Z)$, for almost all sequences $Z_1, Z_2, \ldots$.*

**Proof.** By the Lindeberg central limit theorem, convergence to the given normal limit will occur for every sequence $Z_1, Z_2, \ldots$ for which

$$n^{-1} \sum_{i=1}^n Z_i Z_i^T \to \mathrm{cov}Z$$

and

$$\frac{1}{n} \sum_{i=1}^n \|Z_i\|^2 \mathrm{E}_\xi \xi_i^2 1\{|\xi_i| \times \|Z_i\| > \epsilon\sqrt{n}\} \quad \to \quad 0, \tag{10.1}$$

for all $\epsilon > 0$, where $\mathrm{E}_\xi$ is the conditional expectation given the $Z_1, Z_2, \ldots$. The first condition is true for almost all sequences by the strong law of

large numbers. We now evaluate the second condition. Fix $\epsilon > 0$. Now, for any $\tau > 0$, the sum in (10.1) is bounded by

$$\frac{1}{n}\sum_{i=1}^{n} \|Z_i\|^2 \mathrm{E}[\xi^2 1\{|\xi| > \epsilon/\tau\}] + \frac{1}{n}\sum_{i=1}^{n} \|Z_i\|^2 1\{\|Z_i\| > \sqrt{n}\tau\}.$$

The first sum has an arbitrarily small upper bound in the limit if we choose sufficiently small $\tau$. Since $\mathrm{E}\|Z\|^2 < \infty$, the second sum will go to zero for almost all sequences $Z_1, Z_2, \ldots$. Thus, for almost all sequences $Z_1, Z_2, \ldots$, (10.1) will hold for any $\epsilon > 0$. For, the intersection of the two sets of sequences, all required conditions hold, and the desired result follows.□

**Proof of Theorem 10.4.** Since the processes $\mathbb{G}$, $\mathbb{G}_n$, $\mathbb{G}'_n$ and $\mathbb{G}''_n$ are unaffected if the class $\mathcal{F}$ is replaced with $\{f - Pf : f \in \mathcal{F}\}$, we will assume $\|P\|_{\mathcal{F}} = 0$ throughout the proof, without loss of generality.

(i)⇒(ii): If $\mathcal{F}$ is Donsker, the sequence $\mathbb{G}'_n$ converges in distribution to a Brownian bridge process by the unconditional multiplier central limit theorem (Theorem 10.1). Thus $\mathbb{G}'_n$ is asymptotically measurable. Now, by Lemma 8.17, a Donsker class is totally bounded by the semimetric $\rho_P(f,g) \equiv (P[f-g]^2)^{1/2}$. For each fixed $\delta > 0$ and $f \in \mathcal{F}$, denote $\Pi_\delta f$ to be the closest element in a given, finite $\delta$-net (with respect to the metric $\rho_P$) for $\mathcal{F}$. We have by continuity of the limit process $\mathbb{G}$, that $\mathbb{G} \circ \Pi_\delta \to \mathbb{G}$, almost surely, as $\delta \downarrow 0$. Hence, for any sequence $\delta_n \downarrow 0$,

$$\sup_{h \in BL_1} |\mathrm{E}h(\mathbb{G} \circ \Pi_{\delta_n}) - \mathrm{E}h(\mathbb{G})| \quad \to \quad 0. \tag{10.2}$$

By Lemma 10.5 above, we also have for any fixed $\delta > 0$ that

$$\sup_{h \in BL_1} |\mathrm{E}_\xi h(\mathbb{G}'_n \circ \Pi_\delta) - \mathrm{E}h(\mathbb{G} \circ \Pi_\delta)| \quad \to \quad 0, \tag{10.3}$$

as $n \to \infty$, for almost all sequences $X_1, X_2, \ldots$. To see this, let $f_1, \ldots, f_m$ be the $\delta$-mesh of $\mathcal{F}$ that defines $\Pi_\delta$. Now define the map $A : \mathbb{R}^m \mapsto \ell^\infty(\mathcal{F})$ by $(A(y))(f) = y_k$, where $y = (y_1, \ldots, y_m)$ and the integer $k$ satisfies $\Pi_\delta f = f_k$. Now $h(\mathbb{G} \circ \Pi_\delta) = g(\mathbb{G}(f_1), \ldots, \mathbb{G}(f_m))$ for the function $g : \mathbb{R}^m \mapsto \mathbb{R}$ defined by $g(y) = h(A(y))$. It is not hard to see that if $h$ is bounded Lipschitz on $\ell^\infty(\mathcal{F})$, then $g$ is also bounded Lipschitz on $\mathbb{R}^m$ with a Lipschitz norm no larger than the Lipschitz norm for $h$. Now (10.3) follows from Lemma 10.5. Note also that $BL_1(\mathbb{R}^m)$ is separable with respect to the metric $\rho_{(m)}(f,g) \equiv \sum_{i=1}^{\infty} 2^{-i} \sup_{x \in K_i} |f(x) - g(x)|$, where $K_1 \subset K_2 \subset \cdots$ are compact sets satisfying $\cup_{i=1}^{\infty} K_i = \mathbb{R}^m$. Hence, since $\mathbb{G}'_n \circ \Pi_\delta$ and $\mathbb{G} \circ \Pi_\delta$ are both tight, the supremum in 10.3 can be replaced by a countable supremum. Thus the displayed quantity is measurable, since $h(\mathbb{G}'_n \circ \Pi_\delta)$ is measurable.

Now, still holding $\delta$ fixed,

$$\begin{aligned}\sup_{h\in BL_1}|\mathrm{E}_\xi h(\mathbb{G}'_n\circ\Pi_\delta)-\mathrm{E}_\xi h(\mathbb{G}'_n)| &\le \sup_{h\in BL_1}\mathrm{E}_\xi\,|h(\mathbb{G}'_n\circ\Pi_\delta)-h(\mathbb{G}'_n)|\\ &\le \mathrm{E}_\xi\,\|\mathbb{G}'_n\circ\Pi_\delta-\mathbb{G}'_n\|^*_{\mathcal{F}}\\ &\le \mathrm{E}_\xi\|\mathbb{G}'_n\|^*_{\mathcal{F}_\delta},\end{aligned}$$

where $\mathcal{F}_\delta\equiv\{f-g:\rho_P(f,g)<\delta,f,g\in\mathcal{F}\}$. Thus the outer expectation of the left-hand-side is bounded above by $\mathrm{E}^*\|\mathbb{G}'_n\|_{\mathcal{F}_\delta}$. As we demonstrated in the proof of Theorem 10.1, $\mathrm{E}^*\|\mathbb{G}'_n\|_{\mathcal{F}_{\delta_n}}\to 0$, for any sequence $\delta_n\downarrow 0$. Now, we choose the sequence $\delta_n$ so that it goes to zero slowly enough to ensure that (10.3) still holds with $\delta$ replaced by $\delta_n$. Combining this with (10.2), the desired result follows.

(ii)⇒(i): Let $h(\mathbb{G}'_n)^*$ and $h(\mathbb{G}'_n)_*$ denote measurable majorants and minorants with respect to $(\xi_1,\ldots,\xi_n,X_1,\ldots,X_n)$ jointly. We now have, by the triangle inequality and Fubini's theorem (Lemma 6.14),

$$\begin{aligned}|\mathrm{E}^*h(\mathbb{G}'_n)-\mathrm{E}h(\mathbb{G})| &\le |\mathrm{E}_X\mathrm{E}_\xi h(\mathbb{G}'_n)^*-\mathrm{E}^*_X\mathrm{E}_\xi h(\mathbb{G}'_n)|\\ &\quad+\mathrm{E}^*_X|\mathrm{E}_\xi h(\mathbb{G}'_n)-\mathrm{E}h(\mathbb{G})|,\end{aligned}$$

where $\mathrm{E}_X$ denotes taking the expectation over $X_1,\ldots,X_n$. By (ii) and the dominated convergence theorem, the second term on the right side converges to zero for all $h\in BL_1$. Since the first term on the right is bounded above by $\mathrm{E}_X\mathrm{E}_\xi h(\mathbb{G}'_n)^*-\mathrm{E}_X\mathrm{E}_\xi h(\mathbb{G}'_n)_*$, it also converges to zero since $\mathbb{G}'_n$ is asymptotically measurable. It is easy to see that the same result holds true if $BL_1$ is replaced by the class of all bounded, Lipschitz continuous nonnegative functions $h:\ell^\infty(\mathcal{F})\mapsto\mathbb{R}$, and thus $\mathbb{G}'_n\rightsquigarrow\mathbb{G}$ unconditionally by the Portmanteau theorem (Theorem 7.6). Hence $\mathcal{F}$ is Donsker by the converse part of Theorem 10.1.

(ii)⇒(iii): Since we can assume $\|P\|_{\mathcal{F}}=0$, we have

$$|h(\mathbb{G}'_n)-h(\mathbb{G}''_n)| \le \|\bar{\xi}\mathbb{G}_n\|_{\mathcal{F}}. \tag{10.4}$$

Moreover, since (ii) also implies (i), we have that $\mathrm{E}^*\|\bar{\xi}\mathbb{G}_n\|_{\mathcal{F}}\to 0$ by Lemma 8.17. Thus $\sup_{h\in BL_1}|\mathrm{E}_\xi h(\mathbb{G}'_n)-\mathrm{E}_\xi h(\mathbb{G}''_n)|\to 0$ in outer probability. Since (10.4) also implies that $\mathbb{G}''_n$ is asymptotically measurable, (iii) follows.

(iii)⇒(i): Arguing as we did in the proof that (ii)⇒(i), it is not hard to show that $\mathbb{G}''_n\rightsquigarrow\mathbb{G}$ unconditionally. Now Theorem 10.1 yields that $\mathcal{F}$ is Donsker.□

We now present the outer-almost-sure conditional multiplier central limit theorem:

THEOREM 10.6 *Assume the conditions of Theorem 10.4. Then the following are equivalent:*

*(i) $\mathcal{F}$ is Donsker and $P^*\|f-Pf\|^2_{\mathcal{F}}<\infty$;*

*(ii) $\mathbb{G}'_n\overset{\text{as}^*}{\underset{\xi}{\rightsquigarrow}}\mathbb{G}$ in $\ell^\infty(\mathcal{F})$.*

*(iii)* $\mathbb{G}_n'' \overset{\text{as}*}{\underset{\xi}{\rightsquigarrow}} \mathbb{G}$ *in* $\ell^\infty(\mathcal{F})$.

**Proof.** The equivalence of (i) and (ii) is given in Theorem 2.9.7 of VW, and we omit its proof.

(ii)⇒(iii): As in the proof of Theorem 10.4, we assume that $\|P\|_{\mathcal{F}} = 0$ without loss of generality. Since

$$|h(\mathbb{G}_n') - h(\mathbb{G}_n'')| \quad \leq \quad |\sqrt{n}\bar{\xi}| \times \|\mathbb{P}_n\|_{\mathcal{F}}, \tag{10.5}$$

for any $h \in BL_1$, we have

$$\sup_{h \in BL_1} |\mathrm{E}_\xi h(\mathbb{G}_n') - \mathrm{E}_\xi(\mathbb{G}_n'')| \leq \mathrm{E}_\xi|\sqrt{n}\bar{\xi}| \times \|\mathbb{P}_n\|_{\mathcal{F}} \leq \|\mathbb{P}_n\|_{\mathcal{F}} \overset{\text{as}*}{\to} 0,$$

since the equivalence of (i) and (ii) implies that $\mathcal{F}$ is both Donsker and Glivenko-Cantelli. Hence

$$\sup_{h \in BL_1} |\mathrm{E}_\xi h(\mathbb{G}_n'') - \mathrm{E}h(\mathbb{G})| \overset{\text{as}*}{\to} 0.$$

The relation (10.5) also yields that $\mathrm{E}_\xi h(\mathbb{G}_n'')^* - \mathrm{E}_\xi h(\mathbb{G}_n'')_* \overset{\text{as}*}{\to} 0$, and thus (iii) follows.

(iii)⇒(ii): Let $h \in BL_1$. Since $\mathrm{E}_\xi h(\mathbb{G}_n'')^* - \mathrm{E}h(\mathbb{G}) \overset{\text{as}*}{\to} 0$, we have $\mathrm{E}^* h(\mathbb{G}_n'') \to \mathrm{E}h(\mathbb{G})$. Since this holds for all $h \in BL_1$, we now have that $\mathbb{G}_n'' \rightsquigarrow \mathbb{G}$ unconditionally by the Portmanteau theorem (Theorem 7.6). Now we can invoke Theorem 10.1 to conclude that $\mathcal{F}$ is both Donsker and Glivenko-Cantelli. Now (10.5) implies (ii) by using an argument almost identical to the one used in the previous paragraph.□

### *10.1.3 Bootstrap Central Limit Theorems*

Theorems 10.4 and 10.6 will now be used to prove Theorems 2.6 and 2.7 from Page 20 of Section 2.2.3. Recall that the multinomial bootstrap is obtained by resampling from the data $X_1, \ldots, X_n$, with replacement, $n$ times to obtain a bootstrapped sample $X_1^*, \ldots, X_n^*$. The empirical measure $\hat{\mathbb{P}}_n^*$ of the bootstrapped sample has the same distribution—given the data—as the measure $\hat{\mathbb{P}}_n \equiv n^{-1}\sum_{i=1}^n W_{ni}\delta_{X_i}$, where $W_n \equiv (W_{n1}, \ldots, W_{nn})$ is a multinomial$(n, n^{-1}, \ldots, n^{-1})$ deviate independent of the data. As in Section 2.2.3, let $\hat{\mathbb{P}}_n \equiv n^{-1}\sum_{i=1}^n W_{ni}\delta_{X_i}$ and $\hat{\mathbb{G}}_n \equiv \sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n)$. Also recall the definitions $\tilde{\mathbb{P}}_n \equiv n^{-1}\sum_{i=1}^n (\xi/\bar{\xi})\delta_{X_i}$ and $\tilde{\mathbb{G}}_n \equiv \sqrt{n}(\mu/\tau)(\tilde{\mathbb{P}}_n - \mathbb{P}_n)$, where the weights $\xi_1, \ldots, \xi_n$ are i.i.d. nonnegative, independent of $X_1, \ldots, X_n$, with mean $0 < \mu < \infty$ and variance $0 < \tau^2 < \infty$, and with $\|\xi\|_{2,1} < \infty$. When $\bar{\xi} = 0$, we define $\tilde{\mathbb{P}}_n$ to be zero. Note that the weights $\xi_1, \ldots, \xi_n$ in this section must have $\mu$ subtracted from them and then divided by $\tau$ before they satisfy the criteria of the multiplier weights in the previous section.

**Proof of Theorem 2.6 (Page 20).** The equivalence of (i) and (ii) follows from Theorem 3.6.1 of VW, which proof we omit. We note, however, that a key component of this proof is a clever approximation of the

multinomial weights with i.i.d. Poisson mean 1 weights. We will use this approximation in our proof of Theorem 10.15 below.

We now prove the equivalence of (i) and (iii). Let $\xi_i^\circ \equiv \tau^{-1}(\xi_i - \mu)$, $i = 1, \ldots, n$, and define $\mathbb{G}_n^\circ \equiv n^{-1/2} \sum_{i=1}^n (\xi_i^\circ - \bar{\xi}^\circ)\delta_{X_i}$, where $\bar{\xi}^\circ \equiv n^{-1} \sum_{i=1}^n \xi_i^\circ$. The basic idea is to show the asymptotic equivalence of $\tilde{\mathbb{G}}_n$ and $\mathbb{G}_n^\circ$. Then Theorem 10.4 can be used to establish the desired result. Accordingly,

$$\mathbb{G}_n^\circ - \tilde{\mathbb{G}}_n \quad = \quad \left(1 - \frac{\mu}{\bar{\xi}}\right) \mathbb{G}_n^\circ \quad = \quad \left(\frac{\bar{\xi}}{\mu} - 1\right) \tilde{\mathbb{G}}_n. \tag{10.6}$$

First, assume that $\mathcal{F}$ is Donsker. Since the weights $\xi_1^\circ, \ldots, \xi_n^\circ$ satisfy the conditions of the unconditional multiplier central limit theorem, we have that $\mathbb{G}_n^\circ \rightsquigarrow \mathbb{G}$. Theorem 10.4 also implies that $\mathbb{G}_n^\circ \overset{\mathrm{P}}{\underset{\xi}{\rightsquigarrow}} \mathbb{G}$. Now (10.6) can be applied to verify that $\|\tilde{\mathbb{G}}_n - \mathbb{G}_n^\circ\|_\mathcal{F} \overset{\mathrm{P}}{\to} 0$, and thus $\tilde{\mathbb{G}}_n$ is asymptotically measurable and

$$\sup_{h \in BL_1} \left| \mathrm{E}_\xi h(\mathbb{G}_n^\circ) - \mathrm{E}_\xi h(\tilde{\mathbb{G}}_n) \right| \overset{\mathrm{P}}{\to} 0.$$

Thus (i)⇒(iii).

Second, assume that $\tilde{\mathbb{G}}_n \overset{\mathrm{P}}{\underset{\xi}{\rightsquigarrow}} \mathbb{G}$. It is not hard to show, arguing as we did in the proof of Theorem 10.4 for the implication (ii)⇒(i), that $\tilde{\mathbb{G}}_n \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathcal{F})$ unconditionally. By applying (10.6) again, we now have that $\|\mathbb{G}_n^\circ - \tilde{\mathbb{G}}_n\|_\mathcal{F} \overset{\mathrm{P}}{\to} 0$, and thus $\mathbb{G}_n^\circ \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathcal{F})$ unconditionally. The unconditional multiplier central limit theorem now verifies that $\mathcal{F}$ is Donsker, and thus (iii)⇒(i).□

**Proof of Theorem 2.7 (Page 20).** The equivalence of (i) and (ii) follows from Theorem 3.6.2 of VW, which proof we again omit. We now prove the equivalence of (i) and (iii).

First, assume (i). Then $\mathbb{G}_n^\circ \overset{\mathrm{as}*}{\underset{\xi}{\rightsquigarrow}} \mathbb{G}$ by Theorem 10.6. Fix $\rho > 0$, and note that by using the first equality in (10.6), we have for any $h \in BL_1$ that

$$\left| h(\tilde{\mathbb{G}}_n) - h(\mathbb{G}_n^\circ) \right| \quad \leq \quad 2 \times 1 \left\{ \left| 1 - \frac{\mu}{\bar{\xi}} \right| > \rho \right\} + (\rho \|\mathbb{G}_n^\circ\|_\mathcal{F}) \wedge 1. \tag{10.7}$$

The first term on the right $\overset{\mathrm{as}*}{\to} 0$. Since the map $\|\cdot\|_\mathcal{F} \wedge 1 : \ell^\infty(\mathcal{F}) \mapsto \mathbb{R}$ is in $BL_1$, we have by Theorem 10.6 that $\mathrm{E}_\xi \left[(\rho \|\mathbb{G}_n^\circ\|_\mathcal{F}) \wedge 1\right] \overset{\mathrm{as}*}{\to} \mathrm{E}\left[\|\rho\mathbb{G}\|_\mathcal{F} \wedge 1\right]$. Let the sequence $0 < \rho_n \downarrow 0$ converge slowly enough so that that the first term on the right in (10.7) $\overset{\mathrm{as}*}{\to} 0$ after replacing $\rho$ with $\rho_n$. Since $\mathrm{E}\left[\|\rho_n \mathbb{G}\|_\mathcal{F} \wedge 1\right] \to 0$, we can apply $\mathrm{E}_\xi$ to both sides of (10.7)—after replacing $\rho$ with $\rho_n$—to obtain

$$\sup_{h \in BL_1} \left| h(\tilde{\mathbb{G}}_n) - h(\mathbb{G}_n^\circ) \right| \overset{\mathrm{as}*}{\to} 0.$$

Combining the fact that $h(\mathbb{G}_n^\circ)^* - h(\mathbb{G}_n^\circ)_* \overset{as*}{\to} 0$ with additional applications of (10.7) yields $h(\tilde{\mathbb{G}}_n)^* - h(\tilde{\mathbb{G}}_n)_* \overset{as*}{\to} 0$. Since $h$ was arbitrary, we have established that $\tilde{\mathbb{G}}_n \underset{\xi}{\overset{as*}{\rightsquigarrow}} \mathbb{G}$, and thus (iii) follows.

Second, assume (iii). Fix $\rho > 0$, and note that by using the second equality in (10.6), we have for any $h \in BL_1$ that

$$\left| h(\mathbb{G}_n^\circ) - h(\tilde{\mathbb{G}}_n) \right| \quad \leq \quad 2 \times 1 \left\{ \left| \frac{\bar{\xi}}{\mu} - 1 \right| > \rho \right\} + \left( \rho \|\tilde{\mathbb{G}}_n\|_{\mathcal{F}} \right) \wedge 1.$$

Since the first term on the right $\overset{as*}{\to} 0$, we can use virtually identical arguments to those used in the previous paragraph—but with the roles of $\mathbb{G}_n^\circ$ and $\tilde{\mathbb{G}}_n$ reversed—to obtain that $\mathbb{G}_n^\circ \underset{\xi}{\overset{as*}{\rightsquigarrow}} \mathbb{G}$. Now Theorem 10.6 yields that $\mathcal{F}$ is Donsker, and thus (i) follows.□

### *10.1.4 Continuous Mapping Results*

We now assume a more general set-up, where $\hat{X}_n$ is a bootstrapped process in a Banach space $(\mathbb{D}, \|\cdot\|)$ and is composed of the sample data $\mathcal{X}_n \equiv (X_1, \ldots, X_n)$ and a random weight vector $M_n \in \mathbb{R}^n$ independent of $\mathcal{X}_n$. We do not require that $X_1, \ldots, X_n$ be i.i.d. In this section, we obtain two continuous mapping results. The first result, Proposition 10.7, is a simple continuous mapping results for the very special case of Lipschitz continuous maps. It is applicable to both the in-probability or outer-almost-sure versions of bootstrap consistency. An interesting special case is the map $g(x) = \|x\|$. In this case, the proposition validates the use of the bootstrap to construct asymptotically uniformly valid confidence bands for $\{Pf : f \in \mathcal{F}\}$ whenever $Pf$ is estimated by $\mathbb{P}_n f$ and $\mathcal{F}$ is $P$-Donsker. Now assume that $\hat{X}_n \underset{M}{\overset{P}{\rightsquigarrow}} X$ and that the distribution of $\|X\|$ is continuous. Lemma 10.11 towards the end of this section reveals that $\mathrm{P}(\|\hat{X}_n\| \leq t | \mathcal{X}_n)$ converges uniformly to $P(\|X\| \leq t)$, in probability. A parallel outer almost sure result holds when $\hat{X}_n \underset{M}{\overset{as*}{\rightsquigarrow}} X$.

The second result, Theorem 10.8, is a considerably deeper result for general continuous maps applied to bootstraps which are consistent in probability. Because of this generality, we must require certain measurability conditions on the map $M_n \mapsto \hat{X}_n$. Fortunately, based on the discussion in the paragraph following Theorem 10.4 above, these measurability conditions are easily satisfied when either $\hat{X}_n = \hat{\mathbb{G}}_n$ or $\hat{X}_n = \tilde{\mathbb{G}}_n$. It appears that other continuous mapping results for bootstrapped empirical processes hold, such as for bootstraps which are outer almost surely consistent, but such results seem to be very challenging to verify.

PROPOSITION 10.7 *Let $\mathbb{D}$ and $\mathbb{E}$ be Banach spaces, $X$ a tight random variable on $\mathbb{D}$, and $g : \mathbb{D} \mapsto \mathbb{E}$ Lipschitz continuous. We have the following:*

*(i) If* $\hat{X}_n \overset{\mathrm{P}}{\underset{M}{\rightsquigarrow}} X$, *then* $g(\hat{X}_n) \overset{\mathrm{P}}{\underset{M}{\rightsquigarrow}} g(X)$.

*(ii) If* $\hat{X}_n \overset{\mathrm{as*}}{\underset{M}{\rightsquigarrow}} X$, *then* $g(\hat{X}_n) \overset{\mathrm{as*}}{\underset{M}{\rightsquigarrow}} g(X)$.

**Proof.** Let $c_0 < \infty$ be the Lipschitz constant for $g$, and, without loss of generality, assume $c_0 \geq 1$. Note that for any $h \in BL_1(\mathbb{E})$, the map $x \mapsto h(g(x))$ is an element of $c_0 BL_1(\mathbb{D})$. Thus

$$\begin{aligned} \sup_{h \in BL_1(\mathbb{E})} \left| \mathrm{E}_M h(g(\hat{X}_n)) - \mathrm{E} h(g(X)) \right| &\leq \sup_{h \in c_0 BL_1(\mathbb{D})} \left| \mathrm{E}_M h(\hat{X}_n) - \mathrm{E} h(X) \right| \\ &= c_0 \sup_{h \in BL_1(\mathbb{D})} \left| \mathrm{E}_M h(\hat{X}_n) - \mathrm{E} h(X) \right|, \end{aligned}$$

and the desired result follows by the respective definitions of $\overset{\mathrm{P}}{\underset{M}{\rightsquigarrow}}$ and $\overset{\mathrm{as*}}{\underset{M}{\rightsquigarrow}}$.□

THEOREM 10.8 *Let* $g : \mathbb{D} \mapsto \mathbb{E}$ *be continuous at all points in* $\mathbb{D}_0 \subset \mathbb{D}$, *where* $\mathbb{D}$ *and* $\mathbb{E}$ *are Banach spaces and* $\mathbb{D}_0$ *is closed. Assume that* $M_n \mapsto h(\hat{X}_n)$ *is measurable for every* $h \in C_b(\mathbb{D})$ *outer almost surely. Then if* $\hat{X}_n \overset{\mathrm{P}}{\underset{M}{\rightsquigarrow}} X$ *in* $\mathbb{D}$, *where* $X$ *is tight and* $\mathrm{P}^*(X \in \mathbb{D}_0) = 1$, $g(\hat{X}_n) \overset{\mathrm{P}}{\underset{M}{\rightsquigarrow}} g(X)$.

**Proof.** As in the proof of the implication (ii)⇒(i) of Theorem 10.4, we can argue that $\hat{X}_n \rightsquigarrow X$ unconditionally, and thus $g(\hat{X}_n) \rightsquigarrow g(X)$ unconditionally by the standard continuous mapping theorem. Moreover, we can replace $\mathbb{E}$ with its closed linear span so that the restriction of $g$ to $\mathbb{D}_0$ has an extension $\tilde{g} : \mathbb{D} \mapsto \mathbb{E}$ which is continuous on all of $\mathbb{D}$ by Dugundji's extension theorem (Theorem 10.9 below). Thus $(g(\hat{X}_n), \tilde{g}(\hat{X}_n)) \rightsquigarrow (g(X), \tilde{g}(X))$, and hence $g(\hat{X}_n) - \tilde{g}(\hat{X}_n) \xrightarrow{\mathrm{P}} 0$. Therefore we can assume without loss of generality that $g$ is continuous on all of $\mathbb{D}$. We can also assume without loss of generality that $\mathbb{D}_0$ is a separable Banach space since $X$ is tight. Hence $\mathbb{E}_0 \equiv g(\mathbb{D}_0)$ is also a separable Banach space.

Fix $\epsilon > 0$. There now exists a compact $K \subset \mathbb{E}_0$ such that $\mathrm{P}(g(X) \notin K) < \epsilon$. By Theorem 10.10 below, the proof of which is given in Section 10.4, we know there exists an integer $k < \infty$, elements $z_1, \ldots, z_k \in C[0,1]$, continuous functions $f_1, \ldots, f_k : \mathbb{E} \mapsto \mathbb{R}$, and a Lipschitz continuous function $J : \overline{\mathrm{lin}}\,(z_1, \ldots, z_k) \mapsto \mathbb{E}$, such that the map $x \mapsto T_\epsilon(x) \equiv J\left(\sum_{j=1}^k z_j f_j(x)\right)$ has domain $\mathbb{E}$ and range $\subset \mathbb{E}$ and satisfies $\sup_{x \in K} \|T_\epsilon(x) - x\| < \epsilon$. Let $BL_1 \equiv BL_1(\mathbb{E})$. We now have

$$\begin{aligned} &\sup_{h \in BL_1} \left| \mathrm{E}_M h(g(\hat{X}_n)) - \mathrm{E} h(g(X)) \right| \\ &\quad \leq \sup_{h \in BL_1} \left| \mathrm{E}_M h(T_\epsilon g(\hat{X}_n)) - \mathrm{E} h(T_\epsilon g(X)) \right| \\ &\qquad + \mathrm{E}_M \left\{ \left\| T_\epsilon g(\hat{X}_n) - g(\hat{X}_n) \right\| \wedge 2 \right\} + \mathrm{E} \left\{ \|T_\epsilon g(X) - g(X)\| \wedge 2 \right\}. \end{aligned}$$

However, the outer expectation of the second term on the right converges to the third term, as $n \to \infty$, by the usual continuous mapping theorem. Thus, provided

$$\sup_{h \in BL_1} \left| \mathrm{E}_M h(T_\epsilon g(\hat{X}_n)) - \mathrm{E} h(T_\epsilon g(X)) \right| \quad \stackrel{\mathrm{P}}{\to} \quad 0, \tag{10.8}$$

we have that

$$\begin{aligned} \limsup_{n \to \infty} \mathrm{E}^* & \left\{ \sup_{h \in BL_1} \left| \mathrm{E}_M h(g(\hat{X}_n)) - \mathrm{E} h(g(X)) \right| \right\} \\ & \leq \quad 2\mathrm{E} \left\{ \|T_\epsilon g(X) - g(X)\| \wedge 2 \right\} \\ & \leq \quad 2\mathrm{E} \left\| \{T_\epsilon g(X) - g(X)\} 1\{g(X) \in K\} \right\| + 4\mathrm{P}(g(X) \notin K) \\ & < \quad 6\epsilon. \end{aligned} \tag{10.9}$$

Now note that for each $h \in BL_1$, $h\left(J\left(\sum_{j=1}^k z_j a_j\right)\right) = \tilde{h}(a_1, \ldots, a_k)$ for all $(a_1, \ldots, a_k) \in \mathbb{R}^k$ and some $\tilde{h} \in c_0 BL_1(\mathbb{R}^k)$, where $1 \leq c_0 < \infty$ (this follows since $J$ is Lipschitz continuous and $\left\|\sum_{j=1}^k z_j a_j\right\| \leq \max_{1 \leq j \leq k} |a_j| \times \sum_{j=1}^k \|z_j\|$). Hence

$$\begin{aligned} \sup_{h \in BL_1} & \left| \mathrm{E}_M h(T_\epsilon g(\hat{X}_n)) - \mathrm{E} h(T_\epsilon g(X)) \right| \\ & \leq \quad \sup_{h \in c_0 BL_1(\mathbb{R}^k)} \left| \mathrm{E}_M h(u(\hat{X}_n)) - \mathrm{E} h(u(X)) \right| \\ & = \quad c_0 \sup_{h \in BL_1(\mathbb{R}^k)} \left| \mathrm{E}_M h(u(\hat{X}_n)) - \mathrm{E} h(u(X)) \right|, \end{aligned} \tag{10.10}$$

where $x \mapsto u(x) \equiv (f_1(g(x)), \ldots, f_k(g(x)))$. Fix any $v : \mathbb{R}^k \mapsto [0,1]$ which is Lipschitz continuous (the Lipschitz constant may be $> 1$). Then, since $\hat{X}_n \rightsquigarrow X$ unconditionally, $\mathrm{E}^* \left\{ \mathrm{E}_M v(u(\hat{X}_n))^* - \mathrm{E}_M v(u(\hat{X}_n))_* \right\} \leq \mathrm{E}^* \left\{ v(u(\hat{X}_n))^* - v(u(\hat{X}_n))_* \right\} \to 0$, where sub- and super- script $*$ denote measurable majorants and minorants, respectively, with respect to the joint probability space of $(\mathcal{X}_n, M_n)$. Thus

$$\left| \mathrm{E}_M v(u(\hat{X}_n)) - \mathrm{E}_M v(u(\hat{X}_n))^* \right| \quad \stackrel{\mathrm{P}}{\to} \quad 0. \tag{10.11}$$

Note that we are using at this point the outer almost sure measurability of $M_n \mapsto v(u(\hat{X}_n))$ to ensure that $E_M v(u(\hat{X}_n))$ is well defined, even if the resulting random expectation is not itself measurable.

Now, for every subsequence $n'$, there exists a further subsequence $n''$ such that $\hat{X}_{n''} \stackrel{\mathrm{as}*}{\underset{M}{\rightsquigarrow}} X$. This means that for this subsequence, the set $B$ of data subsequences $\{\mathcal{X}_{n''} : n \geq 1\}$ for which $\mathrm{E}_M v(u(\hat{X}_{n''})) - \mathrm{E} v(u(X)) \to 0$ has inner

probability 1. Combining this with (10.11) and Proposition 7.22, we obtain that $\mathrm{E}_M v(u(\hat{X}_n)) - \mathrm{E}v(u(X)) \xrightarrow{\mathrm{P}} 0$. Since $v$ was an arbitrary real, Lipschitz continuous function on $\mathbb{R}^k$, we now have by Part (i) of Lemma 10.11 below followed by Lemma 10.12 below, that

$$\sup_{h \in BL_1(\mathbb{R}^k)} \left| \mathrm{E}_M h(u(\hat{X}_n)) - \mathrm{E}h(u(X)) \right| \xrightarrow{\mathrm{P}} 0.$$

Combining this with (10.10), we obtain that (10.8) is satisfied. The desired result now follows from (10.9), since $\epsilon > 0$ was arbitrary.□

THEOREM 10.9 *(Dugundji's extension theorem) Let $X$ be an arbitrary metric space, $A$ a closed subset of $X$, $L$ a locally convex linear space (which includes Banach vector spaces), and $f : A \mapsto L$ a continuous map. Then there exists a continuous extension of $f$, $F : X \mapsto L$. Moreover, $F(X)$ lies in the closed linear span of the convex hull of $f(A)$.*

**Proof.** This is Theorem 4.1 of Dugundji (1951), and the proof can be found therein.□

THEOREM 10.10 *Let $\mathbb{E}_0 \subset \mathbb{E}$ be Banach spaces with $\mathbb{E}_0$ separable and $\overline{lin}\,\mathbb{E}_0 \subset \mathbb{E}$. Then for every $\epsilon > 0$ and every compact $K \subset \mathbb{E}_0$, there exists an integer $k < \infty$, elements $z_1, \ldots, z_k \in C[0,1]$, continuous functions $f_1, \ldots, f_k : \mathbb{E} \mapsto \mathbb{R}$, and a Lipschitz continuous function $J : \overline{lin}\,(z_1, \ldots, z_k) \mapsto \mathbb{E}$, such that the map $x \mapsto T_\epsilon(x) \equiv J\left(\sum_{j=1}^k z_j f_j(x)\right)$ has domain $\mathbb{E}$ and range $\subset \mathbb{E}$, is continuous, and satisfies $\sup_{x \in K} \|T_\epsilon(x) - x\| < \epsilon$.*

The proof of this theorem is given in Section 10.4. For the next two lemmas, we use the usual partial ordering on $\mathbb{R}^k$ to define relations between points, e.g., for any $s, t \in \mathbb{R}^k$, $s \le t$ is equivalent to $s_1 \le t_1, \ldots, s_k \le t_k$.

LEMMA 10.11 *Let $X_n$ and $X$ be random variables in $\mathbb{R}^k$ for all $n \ge 1$. Define $\mathcal{S} \subset [\mathbb{R} \cup \{-\infty, \infty\}]^k$ to be the set of all continuity points of $t \mapsto F(t) \equiv \mathrm{P}(X \le t)$ and $H$ to be the set of all Lipschitz continuous functions $h : \mathbb{R}^k \mapsto [0,1]$ (the Lipschitz constants may be $> 1$). Then, provided the expectations are well defined, we have:*

*(i) If $\mathrm{E}[h(X_n)|\mathcal{Y}_n] \xrightarrow{\mathrm{P}} \mathrm{E}h(X)$ for all $h \in H$, then $\sup_{t \in A} |\mathrm{P}(X_n \le t|\mathcal{Y}_n) - F(t)| \xrightarrow{\mathrm{P}} 0$ for all closed $A \subset \mathcal{S}$;*

*(ii) If $\mathrm{E}[h(X_n)|\mathcal{Y}_n] \xrightarrow{\mathrm{as*}} \mathrm{E}h(X)$ for all $h \in H$, then $\sup_{t \in A} |\mathrm{P}(X_n \le t|\mathcal{Y}_n) - F(t)| \xrightarrow{\mathrm{as*}} 0$ for all closed $A \subset \mathcal{S}$.*

**Proof.** Let $t_0 \in \mathcal{S}$. For every $\delta > 0$, there exists $h_1, h_2 \in H$, such that $h_1(u) \le 1\{u \le t_0\} \le h_2(u)$ for all $u \in \mathbb{R}^k$ and $\mathrm{E}[h_2(X) - h_1(X)] < \delta$. Under the condition in (i), we therefore have that $\mathrm{P}(X_n \le t_0|\mathcal{Y}_n) \xrightarrow{\mathrm{P}} F(t_0)$,

since $\delta$ was arbitrary. The conclusion of (i) follows since this convergence holds for all $t_0 \in \mathcal{S}$, since both $\mathrm{P}(X_n \le t|\mathcal{Y}_n)$ and $F(t)$ are monotone in $t$ with range $\subset [0,1]$, and since $[0,1]$ is compact. The proof for Part (ii) follows similarly.□

LEMMA 10.12 *Let $\{F_n\}$ and $F$ be distribution functions on $\mathbb{R}^k$, and let $\mathcal{S} \subset [\mathbb{R} \cup \{-\infty, \infty\}]^k$ be the set of all continuity points of $F$. Then the following are equivalent:*

*(i)* $\sup_{t\in A} |F_n(t) - F(t)| \to 0$ *for all closed* $A \subset \mathcal{S}$.

*(ii)* $\sup_{h\in BL_1(\mathbb{R}^k)} \left|\int_{\mathbb{R}^k} h(dF_n - dF)\right| \to 0$.

The relatively straightforward proof is saved as Exercise 10.5.3.

## 10.2 The Bootstrap for Glivenko-Cantelli Classes

We now present several results for the bootstrap applied to Glivenko-Cantelli classes. The primary use of these results is to assist verification of consistency of bootstrapped estimators. The first theorem (Theorem 10.13) consists of various multiplier bootstrap results, and it is followed by a corollary (Corollary 10.14) which applies to certain weighted bootstrap results. The final theorem of this section (Theorem 10.15) gives corresponding results for the multinomial bootstrap. On a first reading through this section, it might be best to skip the proofs and focus on the results and discussion between the proofs.

THEOREM 10.13 *Let $\mathcal{F}$ be a class of measurable functions, and let $\xi_1, \ldots, \xi_n$ be i.i.d. nonconstant random variables with $0 < \mathrm{E}|\xi| < \infty$ and independent of the sample data $X_1, \ldots, X_n$. Let $\mathbb{W}_n \equiv n^{-1}\sum_{i=1}^n \xi_i(\delta_{X_i} - P)$ and $\tilde{\mathbb{W}}_n \equiv n^{-1}\sum_{i=1}^n (\xi_i - \bar{\xi})\delta_{X_i}$, where $\bar{\xi} \equiv n^{-1}\sum_{i=1}^n \xi_i$. Then the following are equivalent:*

*(i)* $\mathcal{F}$ *is strong Glivenko-Cantelli;*

*(ii)* $\|\mathbb{W}_n\|_{\mathcal{F}} \overset{\mathrm{as}*}{\to} 0$;

*(iii)* $\mathrm{E}_\xi\|\mathbb{W}_n\|_{\mathcal{F}} \overset{\mathrm{as}*}{\to} 0$ *and* $P^*\|f - Pf\|_{\mathcal{F}} < \infty$;

*(iv) For every* $\eta > 0$, $\mathrm{P}\left(\|\mathbb{W}_n\|_{\mathcal{F}} > \eta | \mathcal{X}_n\right) \overset{\mathrm{as}*}{\to} 0$ *and* $P^*\|f - Pf\|_{\mathcal{F}} < \infty$, *where* $\mathcal{X}_n \equiv (X_1, \ldots, X_n)$;

*(v) For every* $\eta > 0$, $\mathrm{P}\left(\|\mathbb{W}_n\|^*_{\mathcal{F}} > \eta | \mathcal{X}_n\right) \overset{\mathrm{as}*}{\to} 0$ *and* $P^*\|f - Pf\|_{\mathcal{F}} < \infty$, *for some version of* $\|\mathbb{W}_n\|^*_{\mathcal{F}}$, *where the superscript* $*$ *denotes a measurable majorant with respect to* $(\xi_1, \ldots, \xi_n, X_1, \ldots, X_n)$ *jointly;*

*(vi)* $\|\tilde{\mathbb{W}}_n\|_{\mathcal{F}} \overset{\mathrm{as}*}{\to} 0$;

*(vii)* $\mathrm{E}_\xi \|\tilde{\mathbb{W}}_n\|_{\mathcal{F}} \overset{\text{as*}}{\to} 0$ *and* $P^*\|f - Pf\|_{\mathcal{F}} < \infty$*;*

*(viii)* *For every* $\eta > 0$, $\mathrm{P}\left( \|\tilde{\mathbb{W}}_n\|_{\mathcal{F}} > \eta \middle| \mathcal{X}_n \right) \overset{\text{as*}}{\to} 0$ *and* $P^*\|f - Pf\|_{\mathcal{F}} < \infty$*;*

*(ix)* *For every* $\eta > 0$, $\mathrm{P}\left( \|\tilde{\mathbb{W}}_n\|^*_{\mathcal{F}} > \eta \middle| \mathcal{X}_n \right) \overset{\text{as*}}{\to} 0$ *and* $P^*\|f - Pf\|_{\mathcal{F}} < \infty$*, for some version of* $\|\tilde{\mathbb{W}}_n\|^*_{\mathcal{F}}$*.*

The lengthy proof is given in Section 4 below. As is shown in the proof, the conditional expectations and conditional probabilities in (iii), (iv), (vii) and (viii) are well defined. This is because the quantities inside of the expectations in Parts (iii) and (vii) (and in the conditional probabilities of (iv) and (viii)) are measurable as functions of $\xi_1, \ldots, \xi_n$ conditional on the data. The distinctions between (iv) and (v) and between (viii) and (ix) are not as trivial as they appear. This is because the measurable majorants involved are computed with respect to $(\xi_1, \ldots, \xi_n, X_1, \ldots, X_1)$ jointly, and thus the differences between $\|\mathbb{W}_n\|_{\mathcal{F}}$ and $\|\mathbb{W}_n\|^*_{\mathcal{F}}$ or between $\|\tilde{\mathbb{W}}_n\|_{\mathcal{F}}$ and $\|\tilde{\mathbb{W}}_n\|_{\mathcal{F}}$ may be nontrivial.

The following corollary applies to a class of weighted bootstraps that includes the Bayesian bootstrap mentioned earlier:

COROLLARY 10.14 *Let* $\mathcal{F}$ *be a class of measurable functions, and let* $\xi_1, \ldots, \xi_n$ *be i.i.d. nonconstant, nonnegative random variables with* $0 < \mathrm{E}\xi < \infty$ *and independent of* $X_1, \ldots, X_n$*. Let* $\tilde{\mathbb{P}}_n \equiv n^{-1} \sum_{i=1}^n (\xi_i/\bar{\xi}) \delta_{X_i}$*, where we set* $\tilde{\mathbb{P}}_n = 0$ *when* $\bar{\xi} = 0$*. Then the following are equivalent:*

*(i)* $\mathcal{F}$ *is strong Glivenko-Cantelli.*

*(ii)* $\|\tilde{\mathbb{P}}_n - \mathbb{P}_n\|_{\mathcal{F}} \overset{\text{as*}}{\to} 0$ *and* $P^*\|f - Pf\|_{\mathcal{F}} < \infty$*.*

*(iii)* $\mathrm{E}_\xi \|\tilde{\mathbb{P}}_n - \mathbb{P}_n\|_{\mathcal{F}} \overset{\text{as*}}{\to} 0$ *and* $P^*\|f - Pf\|_{\mathcal{F}} < \infty$*.*

*(iv)* *For every* $\eta > 0$, $\mathrm{P}\left( \|\tilde{\mathbb{P}}_n - \mathbb{P}_n\|_{\mathcal{F}} > \eta \middle| \mathcal{X}_n \right) \overset{\text{as*}}{\to} 0$ *and* $P^*\|f - Pf\|_{\mathcal{F}} < \infty$*;*

*(v)* *For every* $\eta > 0$, $\mathrm{P}\left( \|\tilde{\mathbb{P}}_n - \mathbb{P}_n\|^*_{\mathcal{F}} > \eta \middle| \mathcal{X}_n \right) \overset{\text{as*}}{\to} 0$ *and* $P^*\|f - Pf\|_{\mathcal{F}} < \infty$*, for some version of* $\|\tilde{\mathbb{P}}_n - \mathbb{P}_n\|^*_{\mathcal{F}}$*.*

*If in addition* $P(\xi = 0) = 0$*, then the requirement that* $P^*\|f - Pf\|_{\mathcal{F}} < \infty$ *in (ii) may be dropped.*

**Proof.** Since the processes $\mathbb{P}_n - P$ and $\tilde{\mathbb{P}}_n - \mathbb{P}_n$ do not change when the class $\mathcal{F}$ is replaced with $\dot{\mathcal{F}} \equiv \{f - Pf : f \in \mathcal{F}\}$, we can assume $\|P\|_{\mathcal{F}} = 0$ without loss of generality. Let the envelope of $\dot{\mathcal{F}}$ be denoted $\dot{F} \equiv \|f\|^*_{\dot{\mathcal{F}}}$. Since multiplying the $\xi_i$ by a constant does not change $\xi_i/\bar{\xi}$, we can also assume $\mathrm{E}\xi = 1$ without loss of generality. The fact that the conditional expressions in (iii) and (iv) are well defined can be argued is in the proof of

Theorem 10.13 (given below in Section 4), and we do not repeat the details here.

(i)⇒(ii): Since

$$\tilde{\mathbb{P}}_n - \mathbb{P}_n - \tilde{\mathbb{W}}_n \quad = \quad \left(\frac{1}{\bar{\xi}} - 1\right) 1\{\bar{\xi} > 0\}\tilde{\mathbb{W}}_n - 1\{\bar{\xi} = 0\}\mathbb{P}_n, \tag{10.12}$$

(ii) follows by Theorem 10.13 and the fact that $\bar{\xi} \stackrel{\text{as*}}{\to} 1$.

(ii)⇒(i): Note that

$$\tilde{\mathbb{P}}_n - \mathbb{P}_n - \tilde{\mathbb{W}}_n \quad = \quad -\left(\bar{\xi} - 1\right) 1\{\bar{\xi} > 0\}(\tilde{\mathbb{P}}_n - \mathbb{P}_n) - 1\{\bar{\xi} = 0\}\mathbb{P}_n. \tag{10.13}$$

The first term on the right $\stackrel{\text{as*}}{\to} 0$ by (ii), while the second term on the right is bounded in absolute value by $1\{\bar{\xi} = 0\}\|\mathbb{P}_n\|_{\dot{\mathcal{F}}} \le 1\{\bar{\xi} = 0\}\mathbb{P}_n\dot{F} \stackrel{\text{as*}}{\to} 0$, by the moment condition.

(ii)⇒(iii): The method of proof will be to use the expansion (10.12) to show that $\mathrm{E}_\xi\|\tilde{\mathbb{P}}_n - \mathbb{P}_n - \tilde{\mathbb{W}}_n\|_{\dot{\mathcal{F}}} \stackrel{\text{as*}}{\to} 0$. Then (iii) will follow by Theorem 10.13 and the established equivalence between (ii) and (i). Along this vein, we have by symmetry followed by an application of Theorem 9.29 that

$$\begin{aligned}
\mathrm{E}_\xi\left\{\left|\frac{1}{\bar{\xi}} - 1\right| 1\{\bar{\xi} > 0\}\|\tilde{\mathbb{W}}_n\|_{\dot{\mathcal{F}}}\right\} \quad &\le \quad \frac{1}{n}\sum_{i=1}^{n} \dot{F}(X_i)\mathrm{E}_\xi\left\{\xi_i\left|\frac{1}{\bar{\xi}} - 1\right| 1\{\bar{\xi} > 0\}\right] \\
&= \quad \mathbb{P}_n\dot{F}\mathrm{E}_\xi\left\{|1 - \bar{\xi}|1\{\bar{\xi} > 0\}\right\} \\
&\stackrel{\text{as*}}{\to} \quad 0.
\end{aligned}$$

Since also $\mathrm{E}_\xi\left[1\{\bar{\xi} = 0\}\right]\|\mathbb{P}_n\|_{\dot{\mathcal{F}}} \stackrel{\text{as*}}{\to} 0$, the desired conclusion follows.

(iii)⇒(iv): This is obvious.

(iv)⇒(i): Consider again expansion (10.13). The moment conditions easily give us, conditional on $X_1, X_2, \ldots$, that $1\{\bar{\xi} = 0\}\|\mathbb{P}_n\|_{\dot{\mathcal{F}}} \le 1\{\bar{\xi} = 0\}\mathbb{P}_n\dot{F} \stackrel{\text{P}}{\to} 0$ for almost all sequences $X_1, X_2, \ldots$. By (iv), we also obtain that $|\bar{\xi} - 1|1\{\bar{\xi} > 0\}\|\tilde{\mathbb{P}}_n - \mathbb{P}_n\|_{\dot{\mathcal{F}}} \stackrel{\text{P}}{\to} 0$ for almost all sequences $X_1, X_2, \ldots$. Thus Assertion (viii) of Theorem 10.13 follows, and we obtain (i).

If $P(\xi = 0) = 0$, then $1\{\bar{\xi} = 0\}\mathbb{P}_n = 0$ almost surely, and we no longer need the moment condition $P\dot{F} < \infty$ in the proofs of (ii)⇒(i) and (ii)⇒(iii), and thus the moment condition in (ii) can be dropped in this setting.

(ii)⇒(v): Assertion (ii) implies that there exists a measurable set $B$ of infinite sequences $(\xi_1, X_1), (\xi_2, X_2), \ldots$ with $\mathrm{P}(B) = 1$ such that $\|\tilde{\mathbb{P}}_n - \mathbb{P}_n\|^*_{\mathcal{F}} \to 0$ on $B$ for some version of $\|\tilde{\mathbb{P}}_n - \mathbb{P}_n\|^*_{\mathcal{F}}$. Let $\mathrm{E}_{\xi,\infty}$ be the expectation taken over the infinite sequence $\xi_1, \xi_2, \ldots$ holding the infinite sequence $X_1, X_2, \ldots$ fixed. By the bounded convergence theorem, we have for any $\eta > 0$ and almost all sequences $X_1, X_2, \ldots$,

$$\begin{aligned}\limsup_{n\to\infty} \mathrm{P}\left(\|\tilde{\mathbb{P}}_n - \mathbb{P}_n\|^*_{\mathcal{F}} > \eta \middle| \mathcal{X}_n\right) &= \limsup_{n\to\infty} \mathrm{E}_{\xi,\infty} 1\left\{\|\tilde{\mathbb{P}}_n - \mathbb{P}_n\|^*_{\mathcal{F}} > \eta\right\} 1\{B\} \\ &= \mathrm{E}_{\xi,\infty} \limsup_{n\to\infty} 1\left\{\|\tilde{\mathbb{P}}_n - \mathbb{P}_n\|^*_{\mathcal{F}} > \eta\right\} 1\{B\} \\ &= 0.\end{aligned}$$

Thus (v) follows.

(v)⇒(iv): This is obvious.□

The following theorem verifies consistency of the multinomial bootstrapped empirical measure defined in Section 10.1.3, which we denote $\hat{\mathbb{P}}_n$, when $\mathcal{F}$ is strong G-C. The proof is given in Section 4 below.

THEOREM 10.15 *Let $\mathcal{F}$ be a class of measurable functions, and let the multinomial vectors $W_n$ in $\hat{\mathbb{P}}_n$ be independent of the data. Then the following are equivalent:*

*(i) $\mathcal{F}$ is strong Glivenko-Cantelli;*

*(ii) $\|\hat{\mathbb{P}}_n - \mathbb{P}_n\|_{\mathcal{F}} \overset{\text{as*}}{\to} 0$ and $P^*\|f - Pf\|_{\mathcal{F}} < \infty$;*

*(iii) $\mathrm{E}_W\|\hat{\mathbb{P}}_n - \mathbb{P}_n\|_{\mathcal{F}} \overset{\text{as*}}{\to} 0$ and $P^*\|f - Pf\|_{\mathcal{F}} < \infty$;*

*(iv) For every $\eta > 0$, $\mathrm{P}\left(\|\hat{\mathbb{P}}_n - \mathbb{P}_n\|_{\mathcal{F}} > \eta \middle| \mathcal{X}_n\right) \overset{\text{as*}}{\to} 0$ and $P^*\|f - Pf\|_{\mathcal{F}} < \infty$;*

*(v) For every $\eta > 0$, $\mathrm{P}\left(\|\hat{\mathbb{P}}_n - \mathbb{P}_n\|^*_{\mathcal{F}} > \eta \middle| \mathcal{X}_n\right) \overset{\text{as*}}{\to} 0$ and $P^*\|f - Pf\|_{\mathcal{F}} < \infty$, for some version of $\|\hat{\mathbb{P}}_n - \mathbb{P}_n\|^*_{\mathcal{F}}$.*

## 10.3 A Simple Z-Estimator Master Theorem

Consider Z-estimation based on the estimating equation $\theta \mapsto \Psi_n(\theta) \equiv \mathbb{P}_n\psi_\theta$, where $\theta \in \Theta \subset \mathbb{R}^p$ and $x \mapsto \psi_\theta(x)$ is a measurable $p$-vector valued function for each $\theta$. This is a special case of the more general Z-estimation approach discussed in Section 2.2.5. Define the map $\theta \mapsto \Psi(\theta) \equiv P\psi_\theta$, and assume $\theta_0 \in \Theta$ satisfies $\Psi(\theta_0) = 0$. Let $\hat{\theta}_n$ be an approximate zero of $\Psi_n$, and let $\hat{\theta}^\circ_n$ be an approximate zero of the bootstrapped estimating equation $\theta \mapsto \Psi^\circ_n(\theta) \equiv \mathbb{P}^\circ_n\psi_\theta$, where $\mathbb{P}^\circ_n$ is either $\tilde{\mathbb{P}}_n$ of Corollary 10.14—with $\xi_1, \ldots, \xi_n$ satisfying the conditions specified in the first paragraph of Section 10.1.3 (the multiplier bootstrap)—or $\hat{\mathbb{P}}_n$ of Theorem 10.15 (the multinomial bootstrap).

The goal of this section is to determine reasonably general conditions under which $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow Z$, where $Z$ is mean zero normally distributed, and $\sqrt{n}(\hat{\theta}^\circ_n - \hat{\theta}_n) \underset{\circ}{\overset{\mathrm{P}}{\rightsquigarrow}} k_0 Z$. Here, we use $\underset{\circ}{\overset{\mathrm{P}}{\rightsquigarrow}}$ to denote either $\underset{\xi}{\overset{\mathrm{P}}{\rightsquigarrow}}$ or $\underset{W}{\overset{\mathrm{P}}{\rightsquigarrow}}$ depending on which bootstrap is being used, and $k_0 = \tau/\mu$ for the multiplier

bootstrap while $k_0 = 1$ for the multinomial bootstrap. One could also estimate the limiting variance rather than use the bootstrap, but there are many settings, such as least absolute deviation regression, where variance estimation may be more awkward than the bootstrap. For theoretical validation of the bootstrap approach, we have the following theorem, which is related to Theorem 2.11 and which utilizes some of the bootstrap results of this chapter:

THEOREM 10.16 *Let $\Theta \subset \mathbb{R}^p$ be open, and assume $\theta_0 \in \Theta$ satisfies $\Psi(\theta_0) = 0$. Also assume the following:*

*(A) For any sequence $\{\theta_n\} \in \Theta$, $\Psi(\theta_n) \to 0$ implies $\|\theta_n - \theta_0\| \to 0$;*

*(B) The class $\{\psi_\theta : \theta \in \Theta\}$ is strong Glivenko-Cantelli;*

*(C) For some $\eta > 0$, the class $\mathcal{F} \equiv \{\psi_\theta : \theta \in \Theta, \|\theta - \theta_0\| \le \eta\}$ is Donsker and $P\|\psi_\theta - \psi_{\theta_0}\|^2 \to 0$ as $\|\theta - \theta_0\| \to 0$;*

*(D) $P\|\psi_{\theta_0}\|^2 < \infty$ and $\Psi(\theta)$ is differentiable at $\theta_0$ with nonsingular derivative matrix $V_{\theta_0}$;*

*(E) $\Psi_n(\hat{\theta}_n) = o_P(n^{-1/2})$ and $\Psi_n^\circ(\hat{\theta}_n^\circ) = o_P(n^{-1/2})$.*

*Then*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow Z \sim N\left(0, V_{\theta_0}^{-1} P[\psi_{\theta_0}\psi_{\theta_0}^T](V_{\theta_0}^{-1})^T\right)$$

*and $\sqrt{n}(\hat{\theta}_n^\circ - \hat{\theta}_n) \overset{\mathrm{P}}{\underset{\circ}{\rightsquigarrow}} k_0 Z$.*

Before giving the proof, we make a few comments about the Conditions (A)–(E) of the theorem. Condition (A) is one of several possible identifiability conditions. Condition (B) is a sufficient condition, when combined with (A), to yield consistency of a zero of $\Psi_n$. This condition is generally reasonable to verify in practice. Condition (C) is needed for asymptotic normality of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ and is also not hard to verify in practice. Condition (D) enables application of the delta method at the appropriate juncture in the proof below, and (E) is a specification of the level of approximation permitted in obtaining the zeros of the estimating equations. See Exercise 10.5.5 below for a specific example of an estimation setting that satisfies these conditions.

**Proof of Theorem 10.16.** By (B) and (E),

$$\|\Psi(\hat{\theta}_n)\| \le \|\Psi_n(\hat{\theta}_n)\| + \sup_{\theta \in \Theta} \|\Psi_n(\theta) - \Psi(\theta)\| \le o_P(1).$$

Thus $\hat{\theta}_n \overset{\mathrm{P}}{\to} \theta_0$ by the identifiability Condition (A). By Assertion (ii) of either Corollary 10.14 or Theorem 10.15 (depending on which bootstrap is used), Condition (B) implies $\sup_{\theta \in \Theta} \|\Psi_n^\circ(\theta) - \Psi(\theta)\| \overset{\mathrm{as}*}{\to} 0$. Thus reapplication of Conditions (A) and (E) yield $\hat{\theta}_n^\circ \overset{\mathrm{P}}{\to} \theta_0$. Note that for the first

part of the proof we will be using unconditional bootstrap results, while the associated conditional bootstrap results will be used only at the end.

By (C) and the consistency of $\hat{\theta}_n$, we have $\mathbb{G}_n\psi_{\hat{\theta}_n} - \mathbb{G}_n\psi_{\theta_0} \xrightarrow{\mathrm{P}} 0$. Since (E) now implies that $\mathbb{G}_n\psi_{\hat{\theta}_n} = \sqrt{n}P(\psi_{\theta_0} - \psi_{\hat{\theta}_n}) + o_P(1)$, we can use the parametric (Euclidean) delta method plus differentiability of $\Psi$ to obtain

$$\sqrt{n}V_{\theta_0}(\theta_0 - \hat{\theta}_n) + \sqrt{n}o_P(\|\hat{\theta}_n - \theta_0\|) \quad = \quad \mathbb{G}_n\psi_{\theta_0} + o_P(1). \tag{10.14}$$

Since $V_{\theta_0}$ is nonsingular, this yields that $\sqrt{n}\|\hat{\theta}_n - \theta_0\|(1 + o_P(1)) = O_P(1)$, and thus $\sqrt{n}(\hat{\theta}_n - \theta_0) = O_P(1)$. Combining this with (10.14), we obtain

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \quad = \quad -V_{\theta_0}^{-1}\sqrt{n}\mathbb{P}_n\psi_{\theta_0} + o_P(1), \tag{10.15}$$

and thus $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow Z$ with the specified covariance.

The first part of Condition (C) and Theorem 2.6 imply that $\mathbb{G}_n^\circ \equiv k_0^{-1}\sqrt{n}(\mathbb{P}_n^\circ - \mathbb{P}_n) \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathcal{F})$ unconditionally, by arguments similar to those used in the (ii)$\Rightarrow$(i) part of the proof of Theorem 10.4. Combining this with the second part of Condition (C), we obtain $k_0\mathbb{G}_n^\circ(\psi_{\hat{\theta}_n^\circ}) + \mathbb{G}_n(\psi_{\hat{\theta}_n^\circ}) - k_0\mathbb{G}_n^\circ(\psi_{\theta_0}) - \mathbb{G}_n(\psi_{\theta_0}) \xrightarrow{\mathrm{P}} 0$. Condition (E) now implies $\sqrt{n}P(\psi_{\theta_0} - \psi_{\hat{\theta}_n^\circ}) = \sqrt{n}\mathbb{P}_n^\circ\psi_{\theta_0} + o_P(1)$. Using similar arguments to those used in the previous paragraph, we obtain

$$\sqrt{n}(\hat{\theta}_n^\circ - \theta_0) = -V_{\theta_0}^{-1}\sqrt{n}\mathbb{P}_n^\circ\psi_{\theta_0} + o_P(1).$$

Combining with (10.15), we have

$$\sqrt{n}(\hat{\theta}_n^\circ - \hat{\theta}_n) = -V_\theta^{-1}\sqrt{n}(\mathbb{P}_n^\circ - \mathbb{P}_n)\psi_{\theta_0} + o_P(1).$$

The desired conditional bootstrap convergence now follows from Theorem 2.6, Part (ii) or Part (iii) (depending on which bootstrap is used).□

## 10.4 Proofs

**Proof of Theorem 10.10.** Fix $\epsilon > 0$ and a compact $K \subset \mathbb{E}_0$. The proof stems from certain properties of separable Banach spaces that can be found in Megginson (1998). Specifically, the fact that every separable Banach space is isometrically isomorphic to a subspace of $C[0,1]$, implies the existence of an isometric isomorphism $J_* : \mathbb{E}_0 \mapsto \mathbb{A}_0$, where $\mathbb{A}_0$ is a subspace of $C[0,1]$. Since $C[0,1]$ has a basis, we know by Theorem 4.1.33 of Megginson (1998) that it also has the "approximation property." This means by Theorem 3.4.32 of Megginson (1998) that since $J_*(K)$ is compact, there exists a finite rank, bounded linear operator $T_* : \mathbb{A}_0 \mapsto \mathbb{A}_0$ such that $\sup_{y \in J_*(K)} \|T_*(y) - y\| < \epsilon$. Because $T_*$ is finite rank, this means there exists elements $z_1, \ldots, z_k \in \mathbb{A}_0 \subset C[0,1]$ and bounded linear functionals

$f_1^*, \ldots, f_k^* : \mathbb{A}_0 \mapsto \mathbb{R}$ such that $T_*(y) = \sum_{j=1}^k z_j f_j^*(y)$. Note that since both $J_*$ and $J_*^{-1} : \mathbb{A}_0 \mapsto \mathbb{E}_0$ are isometric isomorphisms, they are also both Lipschitz continuous with Lipschitz constant 1. This means that the map $x \mapsto \tilde{T}_\epsilon(x) \equiv J_*^{-1}(T_*(J_*(x)))$, with domain and range $\mathbb{E}_0$, satisfies $\sup_{x \in K} \|\tilde{T}_\epsilon(x) - x\| < \infty$.

We now need to verify the existence of several important extensions. By Dugundji's extension theorem (Theorem 10.9 above), there exists a continuous extension of $J_*$, $\tilde{J}_* : \mathbb{E} \mapsto \overline{\text{lin}}\, \mathbb{A}_0$. Also, by the Hahn-Banach extension theorem, there exist bounded linear extensions of $f_1^*, \ldots, f_k^*$, $\tilde{f}_1^*, \ldots, \tilde{f}_k^* : \overline{\text{lin}}\, \mathbb{A}_0 \mapsto \mathbb{R}$. Now let $\tilde{J}$ denote the restriction of $J_*^{-1}$ to the domain $\left\{\sum_{j=1}^k z_j f_j^*(y) : y \in J_*(\mathbb{E}_0)\right\}$. Since $\tilde{J}$ is Lipschitz continuous, as noted previously, we now have by Theorem 10.17 below that there exists a Lipschitz continuous extension of $\tilde{J}$, $J : \overline{\text{lin}}\,(z_1, \ldots, z_k) \mapsto \mathbb{E}$, with Lipschitz constant possibly larger than 1. Now define $x \mapsto f_j(x) \equiv \tilde{f}_j^*(\tilde{J}_*(x))$, for $j = 1, \ldots, k$, and $x \mapsto T_\epsilon(x) \equiv J\left(\sum_{i=1}^k z_j f_j(x)\right)$; and note that $T_\epsilon$ is a continuous extension of $\tilde{T}_\epsilon$. Now the quantities $k$, $z_1, \ldots, z_k$, $f_1, \ldots, f_k$, $J$ and $T_\epsilon$ all satisfy the given requirements.□

**Proof of Theorem 10.13.** Since the processes $\mathbb{P}_n - P$, $\mathbb{W}_n$ and $\tilde{\mathbb{W}}_n$ do not change when $\mathcal{F}$ is replaced by $\dot{\mathcal{F}} \equiv \{f - Pf : f \in \mathcal{F}\}$, we can use for the index set either $\mathcal{F}$ or $\dot{\mathcal{F}} \equiv \{f - Pf : f \in \mathcal{F}\}$ without changing the processes. Define also $\dot{F} \equiv \|f - Pf\|_{\mathcal{F}}^*$. We first need to show that the expectations and probabilities in (iii), (iv), (vii) and (viii) are well defined. Note that for fixed $x_1, \ldots, x_n$ and $a \equiv (a_1, \ldots, a_n) \in \mathbb{R}^n$, the map

$$(a_1, \ldots, a_n) \mapsto \left\| n^{-1} \sum_{i=1}^n a_i f(x_i) \right\|_{\dot{\mathcal{F}}} = \sup_{u \in B} |a^T u|,$$

where $B \equiv \{(f(x_1), \ldots, f(x_n)) : f \in \dot{\mathcal{F}}\} \subset \mathbb{R}^n$. By the continuity of the map $(a, u) \mapsto |a^T u|$ and the separability of $\mathbb{R}$, this map is a measurable function even if the set $B$ is not a Borel set. Thus the conditional expectations and conditional probabilities are indeed well defined.

(i)⇒(ii): Note that $\mathcal{F}$ being $P$-G-C implies that $P\dot{F} < \infty$ by Lemma 8.13. Because $\dot{\mathcal{F}}$ is G-C and $\xi$ is trivially G-C, the desired result follows from Corollary 9.27 (of Section 9.3) and the fact that $\|\xi(f - Pf)\|_{\mathcal{F}}^* \le |\xi| \times \dot{F}$ is integrable.

(ii)⇒(i): Since both $\text{sign}(\xi)$ and $\xi \cdot \dot{\mathcal{F}}$ are $P$-G-C, Corollary 9.27 can be applied to verify that $\text{sign}(\xi) \cdot \xi \cdot \dot{\mathcal{F}} = |\xi| \cdot \dot{\mathcal{F}}$ is also $P$-G-C. We also have by Lemma 8.13 that $P^*\dot{F} < \infty$ since $P|\xi| > 0$. Now we have for fixed $X_1, \ldots, X_n$,

$$(\text{E}\xi)\|\mathbb{P}_n\|_{\dot{\mathcal{F}}} = \|n^{-1} \sum_{i=1}^n (\text{E}\xi_i)\delta_{X_i}\|_{\dot{\mathcal{F}}} \le \text{E}_\xi \|\mathbb{W}_n\|_{\mathcal{F}},$$

and thus $(\mathrm{E}\xi)\mathrm{E}^*\|\mathbb{P}_n - P\|_{\mathcal{F}} \le \mathrm{E}^*\|\mathbb{W}_n\|_{\mathcal{F}}$. By applying Theorem 9.29 twice, we obtain the desired result.

(ii)⇒(iii): Note first that (ii) immediately implies $P|\xi|\dot{F}(X) < \infty$ by Lemma 8.13. Thus $P\dot{F} < \infty$ since $\mathrm{E}|\xi| > 0$. Define $R_n \equiv n^{-1}\sum_{i=1}^n |\xi_i|\dot{F}(X_i)$, and let $B$ be the set of all infinite sequences $(\xi_1, X_1), (\xi_2, X_2), \ldots$ such that $\|\mathbb{W}_n\|_{\mathcal{F}} \overset{\text{as*}}{\to} 0$ and $R_n \to \mathrm{E}[|\xi|\dot{F}]$. Note that the set $B$ has probability 1. Moreover, by the bounded convergence theorem,

$$\begin{aligned}\limsup \mathrm{E}_\xi\|\mathbb{W}_n\|_{\dot{\mathcal{F}}}1\{R_n \le K\} &= \limsup \mathrm{E}_{\xi,\infty}\|\mathbb{W}_n\|_{\dot{\mathcal{F}}}1\{R_n \le K\}1\{B\} \\ &\le \mathrm{E}_{\xi,\infty}\limsup \|\mathbb{W}_n\|^*_{\dot{\mathcal{F}}}1\{R_n \le K\}1\{B\} \\ &= 0,\end{aligned}$$

outer almost surely, for any $K < \infty$. In addition, if we let $S_n = n^{-1}\sum_{i=1}^n |\xi_i|$, we have for any $0 < N < \infty$ that

$$\begin{aligned}\mathrm{E}_\xi\|\mathbb{W}_n\|_{\dot{\mathcal{F}}}1\{R_n > K\} &\le \mathrm{E}_\xi R_n 1\{R_n > K\} \\ &\le \mathrm{E}_\xi\left[S_n 1\{R_n > K\}\right]\mathbb{P}_n\dot{F} \\ &\le \frac{N(\mathrm{E}|\xi|)[\mathbb{P}_n\dot{F}]^2}{K} + \mathrm{E}_\xi[S_n 1\{S_n > N\}]\mathbb{P}_n\dot{F},\end{aligned}$$

where the second-to-last inequality follows by symmetry. By Exercise 10.5.4, the last line of the display has a $\limsup \le N(P\dot{F})^2/K + (\mathrm{E}|\xi|)^2/N$ outer almost surely. Thus, if we let $N = \sqrt{K}$ and allow $K \uparrow \infty$ slowly enough, we ensure that $\limsup_{n\to\infty} \mathrm{E}_\xi\|\mathbb{W}_n\|_{\mathcal{F}} \to 0$, outer almost surely. Hence (iii) follows.

(iii)⇒(iv): This is obvious.

(iv)⇒(ii): (iv) clearly implies that $\|\mathbb{W}_n\|_{\mathcal{F}} \overset{\mathrm{P}}{\to} 0$. Now Lemma 8.16 implies that since the class $|\xi| \times \dot{\mathcal{F}}$ has an integrable envelope, a version of $\|\mathbb{W}_n\|^*_{\mathcal{F}}$ must converge outer almost surely to a constant. Thus (ii) follows.

(ii)⇒(v): Assertion (ii) implies that there exists a measurable set $B$ of infinite sequences $(\xi_1, X_1), (\xi_2, X_2), \ldots$ with $\mathrm{P}(B) = 1$ such that $\|\mathbb{W}_n\|^*_{\mathcal{F}} \to 0$ on $B$ for some version of $\|\mathbb{W}_n\|^*_{\mathcal{F}}$. Thus by the bounded convergence theorem, we have for any $\eta > 0$ and almost all sequences $X_1, X_2, \ldots$,

$$\begin{aligned}\limsup_{n\to\infty} \mathrm{P}\left(\|\mathbb{W}_n\|^*_{\mathcal{F}} > \eta \middle| \mathcal{X}_n\right) &= \limsup_{n\to\infty} \mathrm{E}_{\xi,\infty} 1\left\{\|\mathbb{W}_n\|^*_{\mathcal{F}} > \eta\right\} 1\{B\} \\ &= \mathrm{E}_{\xi,\infty}\limsup_{n\to\infty} 1\left\{\|\mathbb{W}_n\|^*_{\mathcal{F}} > \eta\right\} 1\{B\} \\ &= 0.\end{aligned}$$

Thus (v) follows.

(v)⇒(iv): This is obvious.

(ii)⇒(vi): Note that

$$\|\tilde{\mathbb{W}}_n - \mathbb{W}_n\|_{\dot{\mathcal{F}}} \le |\bar{\xi} - \mathrm{E}\xi| \times |n^{-1}\sum_{i=1}^n \dot{F}(X_i)| + (\mathrm{E}\xi)\|\mathbb{P}_n\|_{\dot{\mathcal{F}}}. \tag{10.16}$$

Since (ii)⇒(i), $P^*\|f - Pf\|_{\mathcal{F}} < \infty$. Thus, since the centered weights $\xi_i - \mathrm{E}\xi$ satisfy the conditions of the theorem as well as the original weights, the right side converges to zero outer almost surely. Hence $\|\tilde{\mathbb{W}}_n\|_{\mathcal{F}} \overset{\text{as*}}{\to} 0$, and (vi) follows.

(vi)⇒(ii): Since $\xi_i$ can be replaced with $\xi_i - \mathrm{E}\xi$ without changing $\tilde{\mathbb{W}}_n$, we will assume without loss of generality that $\mathrm{E}\xi = 0$ (for this paragraph only). Accordingly, (10.16) implies

$$\|\tilde{\mathbb{W}}_n - \mathbb{W}_n\|_{\dot{\mathcal{F}}} \quad \leq \quad |\bar{\xi} - \mathrm{E}\xi| \times |n^{-1} \sum_{i=1}^{n} \dot{F}(X_i)|. \tag{10.17}$$

Thus (ii) will follow by the strong law of large numbers if we can show that $\mathrm{E}\dot{F} < \infty$. Now let $Y_1, \ldots, Y_n$ be i.i.d. $P$ independent of $X_1, \ldots, X_n$, and let $\tilde{\xi}_1, \ldots, \tilde{\xi}_n$ be i.i.d. copies of $\xi_1, \ldots, \xi_n$ independent of $X_1, \ldots, X_n, Y_1, \ldots, Y_n$. Define $\tilde{\mathbb{W}}_{2n} \equiv (2n)^{-1} \sum_{i=1}^{n} \left[(\xi_i - \bar{\xi})f(X_i) + (\tilde{\xi}_i - \bar{\xi})f(Y_i)\right]$ and $\tilde{\mathbb{W}}'_{2n} \equiv (2n)^{-1} \sum_{i=1}^{n} \left[(\tilde{\xi}_i - \bar{\xi})f(X_i) + (\xi_i - \bar{\xi})f(Y_i)\right]$, where $\bar{\xi} \equiv (2n)^{-1} \sum_{i=1}^{n} (\xi_i + \tilde{\xi}_i)$. Since both $\|\tilde{\mathbb{W}}_{2n}\|_{\dot{\mathcal{F}}} \overset{\text{as*}}{\to} 0$ and $\|\tilde{\mathbb{W}}'_{2n}\|_{\dot{\mathcal{F}}} \overset{\text{as*}}{\to} 0$, we have that $\|\tilde{\mathbb{W}}_n - \tilde{\mathbb{W}}'_{2n}\|_{\dot{\mathcal{F}}} \overset{\text{as*}}{\to} 0$. However,

$$\tilde{\mathbb{W}}_n - \tilde{\mathbb{W}}'_{2n} = \frac{1}{n} \sum_{i=1}^{n} \frac{\xi_i - \tilde{\xi}_i}{2} \left[f(X_i) - f(Y_i)\right],$$

and thus $\|n^{-1} \sum_{i=1}^{n} [f(X_i) - f(Y_i)]\|_{\dot{\mathcal{F}}} \overset{\text{as*}}{\to} 0$ by the previously established equivalence between (i) and (ii) and the fact that the new weights $(\xi_i - \tilde{\xi}_i)/2$ satisfy the requisite conditions. Thus

$$\mathrm{E}^*\dot{F} = \mathrm{E}^*\|f(X) - \mathrm{E}f(Y)\|_{\dot{\mathcal{F}}} \leq \mathrm{E}^*\|f(X) - f(Y)\|_{\dot{\mathcal{F}}} < \infty,$$

where the last inequality holds by Lemma 8.13, and (ii) now follows.

(iii)⇒(vii): Since

$$\mathrm{E}_\xi\|\tilde{\mathbb{W}}_n - \mathbb{W}_n\|_{\dot{\mathcal{F}}} \leq (\mathrm{E}|\xi|)\|\mathbb{P}_n\|_{\dot{\mathcal{F}}},$$

we have that $\mathrm{E}_\xi\|\tilde{\mathbb{W}}_n\|_{\dot{\mathcal{F}}} \overset{\text{as*}}{\to} 0$, because (iii) also implies (i).

(vii)⇒(viii): This is obvious.

(viii)⇒(vi): Since $\tilde{\mathbb{W}}_n$ does not change if the $\xi_i$s are replaced by $\xi_i - \mathrm{E}\xi$, we will assume—as we did in the proof that (vi)⇒(ii)—that $\mathrm{E}\xi = 0$ without loss of generality. By reapplication of (10.17) and the strong law of large numbers, we obtain that $\|\tilde{\mathbb{W}}_n\|_{\mathcal{F}} \overset{\text{P}}{\to} 0$. Since the class $|\xi| \times \dot{\mathcal{F}}$ has an integrable envelope, reapplication of Lemma 8.16 yields the desired result.

(vi)⇒(ix): The proof here is identical to the proof that (ii)⇒(v), after exchanging $\mathbb{W}_n$ with $\tilde{\mathbb{W}}_n$.

(ix)⇒(viii): This is obvious.□

**Proof of Theorem 10.15.** The fact that the conditional expressions in Assertions (iii) and (iv) are well defined can be argued as in the proof of Theorem 10.13 above, and we omit the details.

(i)⇒(v): This follows from Lemma 10.18 below since the vectors $W_n/n$ are exchangeable and satisfy the other required conditions.

(v)⇒(iv): This is obvious.

(iv)⇒(i): For each integer $n \geq 1$, generate an infinite sequence of independent random row $n$-vectors $m_n^{(1)}, m_n^{(2)}, \ldots$ as follows. Set $m_1^{(k)} = 1$ for all integers $k \geq 1$, and for each $n > 1$, generate an infinite sequence of i.i.d. Bernoullies $B_n^{(1)}, B_n^{(2)}, \ldots$ with probability of success $1/n$, and set $m_n^{(k)} = [1 - B_n^{(k)}](m_{n-1}^{(k)}, 0) + B_n^{(k)}(0, \ldots, 0, 1)$. Note that for each fixed $n$, $m_n^{(1)}, m_n^{(2)}, \ldots$ are i.i.d. multinomial$(1, n^{-1}, \ldots, n^{-1})$ vectors. Independent of these random quantities, generate an infinite sequence $U_1, U_2, \ldots$ of i.i.d. Poisson random variables with mean 1, and set $N_n = \sum_{i=1}^n U_i$. Also make sure that all of these random quantities are independent of $X_1, X_2, \ldots$. Without loss of generality assume $W_n = \sum_{i=1}^n m_n^{(i)}$ and define $\xi^{(n)} \equiv \sum_{i=1}^{N_n} m_n^{(i)}$. It is easy to verify that the $W_n$ are indeed multinomial$(n, n^{-1}, \ldots, n^{-1})$ vectors as claimed, and that $\xi_i^{(n)}, \ldots, \xi_i^{(n)}$, where $(\xi_1^{(n)}, \ldots, \xi_n^{(n)}) \equiv \xi^{(n)}$, are i.i.d. Possion mean 1 random variables. Note also that these random weights are independent of $X_1, X_2, \ldots$.

Let $\mathbb{W}_n \equiv n^{-1} \sum_{i=1}^n (\xi_i^{(n)} - 1)(\delta_{X_i} - P)$, and note that

$$\hat{\mathbb{P}}_n - \mathbb{P}_n - \mathbb{W}_n = n^{-1} \sum_{i=1}^n (W_{ni} - \xi_i^{(n)})(\delta_{X_i} - P).$$

Since the nonzero elements of $(W_{ni} - \xi_i^{(n)})$ all have the same sign by construction, we have that

$$\begin{aligned} \mathrm{E}_{W,\xi} \|\hat{\mathbb{P}}_n - \mathbb{P}_n - \mathbb{W}_n\|_{\mathcal{F}} &\leq \mathrm{E}_{W,\xi} \|n^{-1} \sum_{i=1}^n |W_{ni} - \xi_i^{(n)}| (\delta_{X_i} - P)\|_{\mathcal{F}} \\ &\leq \left( \mathrm{E} \left| \frac{N_n - n}{n} \right| \right) \left[ \mathbb{P}_n \dot{F} + P\dot{F} \right] \\ &\overset{\text{as*}}{\to} 0, \end{aligned}$$

where the last inequality follows from the exchangeability result $\mathrm{E}_{W,\xi} |W_{ni} - \xi_i^{(n)}| = \mathrm{E}[|N_n - n|/n]$, $1 \leq i \leq n$, and the outer almost sure convergence to zero follows from the fact that $\mathrm{E}[|N_n - n|/n] \leq n^{-1/2}$ combined with the moment conditions. In the forgoing, we have used $\mathrm{E}_{W,\xi}$ to denote taking expectations over $W_n$ and $\xi^{(n)}$ conditional on $X_1, X_2, \ldots$. We have just established that Assertion (iv) holds in Theorem 10.13 with weights $(\xi_1^{(n)} - 1, \ldots, \xi_n^{(n)} - 1)$ that satisfy the necessary conditions. Thus $\mathcal{F}$ is Glivenko-Cantelli.

(v)⇒(ii): Let $\mathrm{E}_{W,\xi,\infty}$ be the expectation over the infinite sequences of the weights conditional on $X_1, X_2, \ldots$. For fixed $\eta > 0$ and $X_1, X_2, \ldots$, we have by the bounded convergence theorem

$$\mathrm{E}_{W,\xi,\infty} \limsup_{n\to\infty} 1\left\{\|\hat{\mathbb{P}}_n - \mathbb{P}_n\|^*_{\mathcal{F}} > \eta\right\} = \limsup_{n\to\infty} \mathrm{E}_{W,\xi,\infty} 1\left\{\|\hat{\mathbb{P}}_n - \mathbb{P}_n\|^*_{\mathcal{F}} > \eta\right\}.$$

But the right side $\to 0$ for almost all $X_1, X_2, \ldots$ by (v). This implies $1\left\{\|\hat{\mathbb{P}}_n - \mathbb{P}_n\|^*_{\mathcal{F}} > \eta\right\} \to 0$, almost surely. Now (ii) follows since $\eta$ was arbitrary.

(ii)⇒(iii): Let $B$ be the set of all sequences of weights and data for which $\|\hat{\mathbb{P}}_n - \mathbb{P}_n\|^*_{\mathcal{F}} \to 0$. From (ii), we know that $B$ is measurable, $P(B) = 1$ and, by the bounded convergence theorem, we have for every $\eta > 0$ and all $X_1, X_2, \ldots$

$$\begin{aligned} 0 &= \mathrm{E}_{W,\xi,\infty} \limsup_{n\to\infty} \left[1\left\{\|\hat{\mathbb{P}}_n - \mathbb{P}_n\|^*_{\mathcal{F}} > \eta\right\} 1\{B\}\right] \\ &= \limsup_{n\to\infty} \mathrm{E}_{W,\xi,\infty} \left[1\left\{\|\hat{\mathbb{P}}_n - \mathbb{P}_n\|^*_{\mathcal{F}} > \eta\right\} 1\{B\}\right]. \end{aligned}$$

Since $P(B) = 1$, this last line implies (v), since $\eta$ was arbitrary, and hence Assertions (i) and (iv) also hold by the previously established equivalences. Fix $0 < K < \infty$, and note that the class $\dot{\mathcal{F}} \cdot 1\{\dot{F} \le K\}$ is strong G-C by Corollary 9.27. Now

$$\begin{aligned} \mathrm{E}_W \|\hat{\mathbb{P}}_n - \mathbb{P}_n\|_{\dot{\mathcal{F}}} &\le \mathrm{E}_W \|\hat{\mathbb{P}}_n - \mathbb{P}_n\|_{\dot{\mathcal{F}}\cdot 1\{\dot{F}\le K\}} \\ &\quad + \mathrm{E}_W \left[(\hat{\mathbb{P}}_n + \mathbb{P}_n)\dot{F} 1\{\dot{F} > K\}\right] \\ &\le \mathrm{E}_W \|\hat{\mathbb{P}}_n - \mathbb{P}_n\|_{\dot{\mathcal{F}}\cdot 1\{\dot{F}\le K\}} + 2\mathbb{P}_n[\dot{F} 1\{\dot{F} > K\}] \\ &\overset{\text{as*}}{\to} 2P[\dot{F} 1\{\dot{F} > K\}], \end{aligned}$$

by Assertion (iv). Since this last term can be made arbitrarily small by choosing $K$ large enough, Assertion (iii) follows.

(iii)⇒(iv): This is obvious.□

THEOREM 10.17 *Let $X, Z$ be metric spaces, with the dimension of $X$ being finite, and let $Y \subset X$. For any Lipschitz continuous map $f : Y \mapsto Z$, there exists a Lipschitz continuous extension $F : X \mapsto Z$.*

**Proof.** This is a simplification of Theorem 2 of Johnson, Lindenstrauss and Schechtman (1986), and the proof can be found therein.□

LEMMA 10.18 *Let $\mathcal{F}$ be a strong Glivenko-Cantelli class of measurable functions. For each $n$, let $(M_{n1}, \ldots, M_{nn})$ be an exchangeable nonnegative random vector independent of $X_1, X_2, \ldots$ such that $\sum_{i=1}^n M_{ni} = 1$ and $\max_{1\le i\le n} |M_{ni}| \overset{\mathrm{P}}{\to} 0$. Then, for every $\eta > 0$,*

$$\mathrm{P}\left( \left\| \sum_{i=1}^{n} M_{ni}(\delta_{X_i} - P) \right\|_{\mathcal{F}}^{*} > \eta \,\middle|\, \mathcal{X}_n \right) \overset{\text{as}*}{\to} 0.$$

This is Lemma 3.6.16 of VW, and we omit the proof.

## 10.5 Exercises

**10.5.1.** Show the following:

(a) $\|\cdot\|_{2,1}$ is a norm on the space of real, square-integrable random variables.

(b) For any real random variable $\xi$, and any $r > 2$, $(1/2)\|\xi\|_2 \le \|\xi\|_{2,1} \le (r/(r-2))\|\xi\|_r$. Hints: For the first inequality, show that $\mathrm{E}|\xi|^2 \le 2\int_0^\infty \mathrm{P}(|\xi| > u)u du \le 2\|\xi\|_{2,1} \times \|\xi\|_2$. For the second inequality, show first that

$$\|\xi\|_{2,1} \le a + \int_a^\infty \left( \frac{\|\xi\|_r^r}{x^r} \right)^{1/2} dx$$

for any $a > 0$.

**10.5.2.** Show that for any $p > 1$, and any real i.i.d. $X_1, \ldots, X_n$ with $\mathrm{E}|X|^p < \infty$, we have

$$\mathrm{E} \max_{1 \le i \le n} \frac{|X_i|}{n^{1/p}} \to 0,$$

as $n \to \infty$. Hint: Show first that for any $x > 0$,

$$\limsup_{n\to\infty} \mathrm{P}\left( \max_{1 \le i \le n} \frac{|X_i|}{n^{1/p}} > x \right) \le 1 - \exp\left( -\frac{\mathrm{E}|X|^p}{x^p} \right).$$

**10.5.3.** Prove Lemma 10.12.

**10.5.4.** Let $\xi_1, \ldots, \xi_n$ be i.i.d. nonnegative random variables with $\mathrm{E}\xi < \infty$, and denote $S_n = n^{-1}\sum_{i=1}^n \xi_i$. Show that

$$\lim_{m\to\infty} \limsup_{n\to\infty} \mathrm{E}[S_n 1\{S_n > m\}] = 0.$$

Hint: Theorem 9.29 may be useful here.

**10.5.5.** Assume that, given the covariate $Z \in \mathbb{R}^p$, $Y$ is Bernoulli with probability of success $e^{\theta^T Z}/(1 + e^{\theta^T Z})$, where $\theta \in \Theta = \mathbb{R}^p$ and $\mathrm{E}[ZZ^T]$ is positive definite. Assume that we observe an i.i.d. sample $(Y_1, Z_1), \ldots, (Y_n, Z_n)$ generated from this model with true parameter $\theta_0 \in \mathbb{R}$. Show that the conditions of Theorem 10.16 are satisfied for Z-estimators based on

$$\psi_\theta(y, z) = Z\left( Y - \frac{e^{\theta^T Z}}{1 + e^{\theta^T Z}} \right).$$

Note that one of the challenges here is the noncompactness of $\Theta$.

## 10.6 Notes

Much of the material in Section 10.1 is inspired by Chapters 2.9 and 3.6 of VW, although the results for the weights $(\xi_1 - \bar{\xi}, \ldots, \xi_n - \bar{\xi})$ and $(\xi_1/\bar{\xi} - 1, \ldots, \xi_n/\bar{\xi})$ and the continuous mapping results are essentially new. The equivalence of Assertions (i) and (ii) of Theorem 10.4 is Theorem 2.9.2 of VW, while the equivalence of (i) and (ii) of Theorem 10.6 is Theorem 2.9.6 of VW. Lemma 10.5 is Lemma 2.9.5 of VW. Theorem 10.16 is an expansion of Theorem 5.21 of van der Vaart (1998), and Part (b) of Exercise 10.5.1 is Exercise 2.9.1 of VW.

# 11
# Additional Empirical Process Results

In this chapter, we study several additional empirical process results that are useful but don't fall neatly into the framework of the other chapters. Because the contents of this chapter are somewhat specialized, some readers may want to skip it the first time they read the book. Although some of the results given herein will be used in later chapters, the results of this chapter are not really necessary for a philosophical understanding of the remainder of the book. On the other hand, this chapter does contain results and references that are useful for readers interested in the deeper potential of empirical process methods in genuinely hard statistical problems.

We first discuss bounding tail probabilities and moments of $\|\mathbb{G}_n\|_{\mathcal{F}}$. These results will be useful in Chapter 14 for determining rates of convergence of M-estimators. We then discuss Donsker results for classes composed of sequences of functions and present several related statistical applications. After this, we discuss contiguous alternative probability measures $P_n$ that get progressively closer to a fixed "null" probability measure $P$ as $n$ gets larger. These results will be useful in Part III of the book, especially in Chapter 18, where we discuss optimality of tests.

We then discuss weak convergence of sums of independent but not identically distributed stochastic processes which arise, for example, in clinical trials with non-independent randomization schemes such as biased coin designs (see, for example, Wei, 1978). We develop this topic in some depth, discussing both a central limit theorem and validity of a certain weighted bootstrap procedure. We also specialize these results to empirical processes based on i.i.d. data but with functions classes $\mathcal{F}_n$ that change with $n$.

The final topic we cover is Donsker results for dependent observations. Here, our brief treatment is primarily meant to introduce the subject and point the interested reader toward the key results and references.

## 11.1 Bounding Moments and Tail Probabilities

We first consider bounding moments of $\|\mathbb{G}_n\|^*_{\mathcal{F}}$ under assumptions similar to the Donsker theorems of Chapter 8. While these do not provide sharp bounds, the bounds are still useful for certain problems. We need to introduce some slightly modified entropy integrals. The first is based on a modified uniform entropy integral:

$$J^*(\delta, \mathcal{F}) \equiv \sup_Q \int_0^\delta \sqrt{1 + \log N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))} d\epsilon,$$

where the supremum is over all finitely discrete probably measures $Q$ with $\|F\|_{Q,2} > 0$. The only difference between this and the previously defined uniform entropy integral $J(\delta, \mathcal{F}, L_2)$, is the presence of the 1 under the radical. The following theorem, which we give without proof, is a subset of Theorem 2.14.1 of VW:

THEOREM 11.1 *Let $\mathcal{F}$ be a $P$-measurable class of measurable functions, with measurable envelope $F$. Then, for each $p \geq 1$,*

$$\| \|\mathbb{G}_n\|^*_{\mathcal{F}} \|_{P,p} \leq c_p J^*(1, \mathcal{F}) \|F\|_{P,2 \wedge p},$$

*where the constant $c_p < \infty$ depends only on $p$.*

We next provide an analogue of Theorem 11.1 for bracketing entropy, based on the modified bracketing integral:

$$J^*_{[]}(\delta, \mathcal{F}) \equiv \int_0^\delta \sqrt{1 + \log N_{[]}(\epsilon \|F\|_{P,2}, \mathcal{F}, L_2(P))} d\epsilon.$$

The difference between this definition and the previously defined $J_{[]}(\delta, \mathcal{F}, L_2(P))$ is twofold: the presence of the 1 under the radical and a rescaling of $\epsilon$ by the factor $\|F\|_{P,2}$. The following theorem, a subset of Theorem 2.14.2 of VW, is given without proof:

THEOREM 11.2 *Let $\mathcal{F}$ be a class of measurable functions with measurable envelope $F$. Then*

$$\| \|\mathbb{G}_n\|^*_{\mathcal{F}} \|_{P,1} \leq c J^*_{[]}(1, \mathcal{F}) \|F\|_{P,2},$$

*for some universal constant $c < \infty$.*

We now provide an additional moment bound based on another modification of the bracketing integral. This modification is

$$\tilde{J}_{[]}(\delta, \mathcal{F}, \|\cdot\|) \equiv \int_0^\delta \sqrt{1 + \log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)}\, d\epsilon.$$

Note that unlike the definition of $J^*_{[]}$ above, the size of the brackets used to compute $N_{[]}$ is not standardized by the norm of the envelope $F$. The following theorem, which we give without proof, is Lemma 3.4.2 of VW:

THEOREM 11.3 *Let $\mathcal{F}$ be a class of measurable functions such that $Pf^2 < \delta^2$ and $\|f\|_\infty \le M$ for every $f \in \mathcal{F}$. Then*

$$\| \|\mathbb{G}_n\|^*_{\mathcal{F}} \|_{P,1} \le c\tilde{J}_{[]}(\delta, \mathcal{F}, L_2(P)) \left[1 + \frac{\tilde{J}_{[]}(\delta, \mathcal{F}, L_2(P))}{\delta^2\sqrt{n}} M\right],$$

*for some universal constant $c < \infty$.*

For some problems, the previous moment bounds are sufficient. In other settings, more refined tail probability bounds are needed. To accomplish this, stronger assumptions are needed for the involved function classes. Recall the definition of pointwise measurable (PM) classes from Section 8.2. The following tail probability results, the proofs of which are given in Section 11.7 below, apply only to bounded and PM classes:

THEOREM 11.4 *Let $\mathcal{F}$ be a pointwise measurable class of functions $f : \mathcal{X} \mapsto [-M, M]$, for some $M < \infty$, such that*

$$\sup_Q \log N\left(\epsilon, \mathcal{F}, L_2(Q)\right) \le K\left(\frac{1}{\epsilon}\right)^W, \tag{11.1}$$

*for all $0 < \epsilon \le M$ and some constants $0 < W < 2$ and $K < \infty$, where the supremum is taken over all finitely discrete probability measures. Then*

$$\| \|\mathbb{G}_n\|^*_{\mathcal{F}} \|_{\psi_2} \le c,$$

*for all $n \ge 1$, where $c < \infty$ depends only on $K$, $W$ and $M$.*

Examples of interesting function classes that satisfy the conditions of Theorem 11.4 are bounded VC classes that are also PM, and the set of all non-decreasing distribution functions on $\mathbb{R}$. This follows from Theorem 9.3 and Lemma 9.11, since the class of distribution functions can be shown to be PM. To see this last claim, for each integer $m \ge 1$, let $\mathcal{G}_m$ be the class of empirical distribution functions based on a sample of size $m$ from the rationals union $\{-\infty, \infty\}$. It is not difficult to show that $\mathcal{G} \equiv \cup_{m\ge1}\mathcal{G}_m$ is countable and that for each distribution function $f$, there exists a sequence $\{g_m\} \in \mathcal{G}$ such that $g_m(x) \to f(x)$, as $m \to \infty$, for each $x \in \mathbb{R}$.

THEOREM 11.5 *Let $\mathcal{F}$ be a pointwise measurable class of functions $f : \mathcal{X} \mapsto [-M, M]$, for some $M < \infty$, such that*

$$N_{[]}\left(\epsilon, \mathcal{F}, L_2(P)\right) \leq K\left(\frac{1}{\epsilon}\right)^W, \tag{11.2}$$

*for all $0 < \epsilon \leq M$ and some constants $K, W < \infty$. Then*

$$\| \, \|\mathbb{G}_n\|_{\mathcal{F}}^* \|_{\psi_2} \leq c,$$

*for all $n \geq 1$, where $c < \infty$ depends only on $K$, $W$ and $M$.*

An example of a function class that satisfies the conditions of Theorem 11.5, are the Lipschitz classes of Theorem 9.23 which satisfy Condition 9.4, provided $T$ is separable and $N(\epsilon, T, d) \leq K(1/\epsilon)^W$ for some constants $K, W < \infty$. This will certainly be true if $(T, d)$ is a Euclidean space.

By Lemma 8.1, if the real random variable $X$ satisfies $\|X\|_{\psi_2} < \infty$, then the tail probabilities of $X$ are "subgaussian" in the sense that $P(|X| > x) \leq Ke^{-Cx^2}$ for some constants $K < \infty$ and $C > 0$. These results can be significantly refined under stronger conditions to yield more precise bounds on the constants. Some results along this line can be found in Chapter 2.14 of VW. A very strong result applies to the empirical distribution function $\mathbb{F}_n$, where $\mathcal{F}$ consists of left half-lines in $\mathbb{R}$:

THEOREM 11.6 *For any i.i.d. sample $X_1, \ldots, X_n$ with distribution $F$,*

$$\mathrm{P}\left(\sup_{t \in \mathbb{R}} \sqrt{n}|\mathbb{F}_n(t) - F(t)| > x\right) \leq 2e^{-2x^2},$$

*for all $x > 0$.*

**Proof.** This is the celebrated result of Dvoretzky, Kiefer and Wolfowitz (1956), given in their Lemma 2, as refined by Massart (1990) in his Corollary 1. We omit the proof of their result but note that their result applies to the special case where $F$ is continuous. We now show that it also applies when $F$ may be discontinuous. Without loss of generality, assume that $F$ has discontinuities, and let $t_1, \ldots, t_m$ be the locations of the discontinuities of $F$, where $m$ may be infinity. Note that the number of discontinuities can be at most countable. Let $p_1, \ldots, p_m$ be the jump sizes of $F$ at $t_1, \ldots, t_m$. Now let $U_1, \ldots, U_n$ be i.i.d. uniform random variables independent of the $X_1, \ldots, X_n$, and define new random variables

$$Y_i = X_i + \sum_{j=1}^{m} p_j \left[1\{X_i > t_j\} + U_i 1\{X_i = t_j\}\right],$$

$1 \leq i \leq n$. Define also the transformation $t \mapsto T(t) = t + \sum_{j=1}^m p_j 1\{t \geq t_j\}$; let $\mathbb{F}_n^*$ be the empirical distribution of $Y_1, \ldots, Y_n$; and let $F^*$ be the distribution of $Y_1$. It is not hard to verify that

$$\begin{aligned}\sup_{t\in\mathbb{R}}|\mathbb{F}_n(t)-F(t)| &= \sup_{t\in\mathbb{R}}|\mathbb{F}_n^*(T(t))-F^*(T(t))| \\ &\leq \sup_{s\in\mathbb{R}}|\mathbb{F}_n^*(s)-F^*(s)|,\end{aligned}$$

and the desired result now follows since $F^*$ is continuous.□

## 11.2 Sequences of Functions

Whether a countable class of functions $\mathcal{F}$ is $P$-Donsker can be verified using the methods of Chapters 9 and 10, but sometimes the special structure of certain countable classes simplifies the evaluation. This is true for certain classes composed of sequences of functions. The following is our first result in this direction:

THEOREM 11.7 *Let $\{f_i, i \geq 1\}$ be any sequence of measurable functions satisfying $\sum_{i=1}^{\infty} P(f_i - Pf_i)^2 < \infty$. Then the class*

$$\left\{\sum_{i=1}^{\infty} c_i f_i : \sum_{i=1}^{\infty}|c_i| \leq 1 \text{ and the series converges pointwise}\right\}$$

*is $P$-Donsker.*

**Proof.** Since the class given in the conclusion of the theorem is the pointwise closure of the symmetric convex hull (see the comments given in Section 9.1.1 just before Theorem 9.4) of the class $\{f_i\}$, it is enough to verify that $\{f_i\}$ is Donsker, by Theorem 9.30. To this end, fix $\epsilon > 0$ and define for each positive integer $m$, a partition $\{f_i\} = \cup_{i=1}^{m+1}\mathcal{F}_i$ as follows. For each $i = 1, \ldots, m$, let $\mathcal{F}_i$ consist of the single point $f_i$, and let $\mathcal{F}_{m+1} = \{f_{m+1}, f_{m+2}, \ldots\}$. Since $\sup_{f,g\in\mathcal{F}_i}|\mathbb{G}_n(f-g)| = 0$ (trivially) for $i = 1, \ldots, m$, we have, by Chebyshev's inequality,

$$\begin{aligned}\mathrm{P}\left(\sup_i \sup_{f,g\in\mathcal{F}_i}|\mathbb{G}_n(f-g)| > \epsilon\right) &\leq \mathrm{P}\left(\sup_{f\in\mathcal{F}_{m+1}}|\mathbb{G}_n f| > \frac{\epsilon}{2}\right) \\ &\leq \frac{4}{\epsilon^2}\sum_{i=m+1}^{\infty} P(f_i - Pf_i)^2.\end{aligned}$$

Since this last term can be made arbitrarily small by choosing $m$ large enough, and since $\epsilon$ was arbitrary, the desired result now follows by Theorem 2.1 via Lemma 7.20.□

When the sequence $\{f_i\}$ satisfies $Pf_if_j = 0$ whenever $i \neq j$, Theorem 11.7 can be strengthened as follows:

THEOREM 11.8 *Let $\{f_i, i \geq 1\}$ be any sequence of measurable functions satisfying $Pf_if_j = 0$ for all $i \neq j$ and $\sum_{i=1}^{\infty} Pf_i^2 < \infty$. Then the class*

$$\left\{\sum_{i=1}^{\infty} c_i f_i : \textstyle\sum_{i=1}^{\infty} c_i^2 \le 1 \textit{ and the series converges pointwise}\right\}$$

*is P-Donsker.*

**Proof.** Since the conditions on $c \equiv (c_1, c_2, \ldots)$ ensure $\sum_{i=1}^{\infty} c_i f_i \le \sqrt{\sum_{i=1}^{\infty} f_i^2}$, we have by the dominated convergence theorem that pointwise converging sums also converge in $L_2(P)$. Now we argue that the class $\mathcal{F}$ of all of these sequences is totally bounded in $L_2(P)$. This follows because $\mathcal{F}$ can be arbitrarily closely approximated by a finite-dimensional set, since

$$P\left(\sum_{i>m} c_i f_i\right)^2 = \sum_{i>m} c_i^2 P f_i^2 \le \sum_{i>m} P f_i^2 \to 0,$$

as $m \to \infty$. Thus the theorem is proved if we can show that the sequence $\mathbb{G}_n$, as a process indexed by $\mathcal{F}$, is asymptotically equicontinuous with respect to the $L_2(P)$-seminorm. Accordingly, note that for any $f = \sum_{i=}^{\infty} c_i f_i$, $g = \sum_{i=1}^{\infty} d_i f_i$, and integer $k \ge 1$,

$$\begin{aligned}|\mathbb{G}_n(f) - \mathbb{G}_n(g)|^2 &= \left|\sum_{i=1}^{\infty}(c_i - d_i)\mathbb{G}_n(f_i)\right|^2 \\ &\le 2\sum_{i=1}^{k}(c_i - d_i)^2 P f_i^2 \sum_{i=1}^{k} \frac{\mathbb{G}_n^2(f_i)}{P f_i^2} + 2\sum_{i=k+1}^{\infty}(c_i - d_i)^2 \sum_{i=k+1}^{\infty} \mathbb{G}_n^2(f_i).\end{aligned}$$

Since, by assumption, $\|c - d\| \le \|c\| + \|d\| \le 2$ (here, $\|\cdot\|$ is the infinite-dimensional Euclidean norm), the above expression is bounded by

$$2\|f - g\|_{P,2}^2 \sum_{i=1}^{k} \frac{\mathbb{G}_n^2(f_i)}{P f_i^2} + 8\sum_{i=k+1}^{\infty} \mathbb{G}_n^2(f_i).$$

Now take the supremum over all pairs of series $f$ and $g$ with $\|f - g\|_{P,2} < \delta$. Since also $E\mathbb{G}_n^2(f_i) \le P f_i^2$, the expectation is bounded above by $2\delta^2 k + 8\sum_{i=k+1}^{\infty} P f_i^2$. This quantity can now be made arbitrarily small by first choosing $k$ large and then choosing $\delta$ small enough.□

If the functions $\{f_i\}$ involved in the preceding theorem are an orthonormal sequence $\{\psi_i\}$ in $L_2(P)$, then the result can be reexpressed in terms of an *elliptical class* for a fixed sequence of constants $\{b_i\}$:

$$\mathcal{F} \equiv \left\{\sum_{i=1}^{\infty} c_i \psi_i : \sum_{i=1}^{\infty} \frac{c_i^2}{b_i^2} \le 1 \text{ and the series converges pointwise}\right\}.$$

More precisely, Theorem 11.8 implies that $\mathcal{F}$ is $P$-Donsker if $\sum_{i=1}^{\infty} b_i^2 < \infty$. Note that a sufficient condition for the stated pointwise convergence to hold

at the point $x$ for all $\{c_i\}$ satisfying $\sum_{i=1}^{\infty} c_i^2/b_i^2 \leq 1$ is for $\sum_{i=1}^{\infty} b_i^2\psi_i^2(x) < \infty$. A very important property of an empirical process indexed by an elliptical class $\mathcal{F}$ is the following:

$$\|\mathbb{G}_n\|_{\mathcal{F}}^2 = \sup_{f\in\mathcal{F}} \left| \sum_{i=1}^{\infty} c_i \mathbb{G}_n(\psi_i) \right|^2 = \sum_{i=1}^{\infty} b_i^2 \mathbb{G}_n^2(\psi_i). \tag{11.3}$$

In the central quantity, each function $f \in \mathcal{F}$ is represented by its series representation $\{c_i\}$. For the second equality, it is easy to see that the last term is an upper bound for the second term by the Cauchy-Schwartz inequality combined with the fact that $\sum_{i=1}^{\infty} c_i^2/b_i^2 \leq 1$. The next thing to note is that this maximum can be achieved by setting $c_i = b_i^2\mathbb{G}_n(\psi_i)/\sqrt{\sum_{i=1}^{\infty} b_i^2\mathbb{G}_n^2(\psi)}$.

An important use for elliptical classes is to characterize the limiting distribution of one- and two- sample Cramér-von Mises, Anderson-Darling, and Watson statistics. We will now demonstrate this for both the Cramér-von Mises and Anderson-Darling statistics. For a study of the one-sample Watson statistic, see Example 2.13.4 of VW. We now have the following key result, the proof of which we give in Section 11.7. Note that the result (11.3) plays an important role in the proof.

THEOREM 11.9 *Let $\mathbb{P}_n$ be the empirical distribution of an i.i.d. sample of uniform $[0,1]$ random variables, let $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$ be the classical empirical process indexed by $t \in [0,1]$, and let $Z_1, Z_2, \ldots$ be an i.i.d. sequence of standard normal deviates independent of $\mathbb{P}_n$. Also define the function classes*

$$\mathcal{F}_1 \equiv \left\{ \sum_{j=1}^{\infty} c_j\sqrt{2}\cos\pi jt : \sum_{j=1}^{\infty} c_j^2\pi^2 j^2 \leq 1 \right\}$$

*and*

$$\mathcal{F}_2 \equiv \left\{ \sum_{j=1}^{\infty} c_j\sqrt{2}p_j(2t-1) : \sum_{j=1}^{\infty} c_j^2 j(j+1) \leq 1 \text{ and pointwise convergence} \right\},$$

*where the functions $p_0(u) \equiv (1/2)\sqrt{2}$, $p_1(u) \equiv (1/2)\sqrt{6}u$, $p_2(u) \equiv (1/4)\sqrt{10}$ $\times(3u^2-1)$, $p_3(u) \equiv (1/4)\sqrt{14}(5u^3-3u)$, and so on, are the orthonormalized Legendre polynomials in $L_2[-1,1]$. Then the following are true:*

*(i) The one-sample Cramér-von Mises statistic for uniform data satisfies*

$$\int_0^1 \mathbb{G}_n^2(t)dt = \|\mathbb{G}_n\|_{\mathcal{F}_1}^2 \rightsquigarrow \frac{1}{\pi^2}\sum_{j=1}^{\infty}\frac{Z_j^2}{j^2} \equiv T_1.$$

*(ii) The one-sample Anderson-Darling statistic for uniform data satisfies*

$$\int_0^1 \frac{\mathbb{G}_n^2(t)}{t(1-t)}dt = \|\mathbb{G}_n\|_{\mathcal{F}_2}^2 \rightsquigarrow \sum_{j=1}^{\infty} \frac{Z_j^2}{j(j+1)} \equiv T_2.$$

Theorem 11.9 applies to testing whether i.i.d. real data $X_1, \ldots, X_n$ comes from an arbitrary continuous distribution $F$. This is realized by replacing $t$ with $F(x)$ throughout the theorem. A more interesting result can be obtained by applying the theorem to testing whether two samples have the same distribution. Let $\hat{F}_{n,j}$ be the empirical distribution of sample $j$ of $n_j$ i.i.d. real random variables, $j = 1, 2$, where the two samples are independent, and where $n = n_1 + n_2$. Let $\hat{F}_{n,0} \equiv (n_1\hat{F}_{n,1} + n_2\hat{F}_{n,2})/n$ be the pooled empirical distribution. The two-sample Cramér-von Mises statistic is

$$\hat{T}_1 \equiv \frac{n_1 n_2}{n_1 + n_1} \int_{-\infty}^{\infty} \left(\hat{F}_{n,1}(s) - \hat{F}_{n,2}(s)\right)^2 d\hat{F}_{n,0}(s),$$

while the two-sample Anderson-Darling statistics is

$$\hat{T}_2 \equiv \frac{n_1 n_2}{n_1 + n_1} \int_{-\infty}^{\infty} \frac{\left(\hat{F}_{n,1}(s) - \hat{F}_{n,2}(s)\right)^2}{\hat{F}_{n,0}(s)\left[1 - \hat{F}_{n,0}(s)\right]} d\hat{F}_{n,0}(s).$$

The proof of the following corollary is given in Section 11.7:

COROLLARY 11.10 *Under the null hypothesis that the two samples come from the same continuous distribution $F_0$, $\hat{T}_j \rightsquigarrow T_j$, as $n_1 \wedge n_2 \rightsquigarrow \infty$, for $j = 1, 2$.*

Since the limiting distributions do not depend on $F_0$, critical values can be easily calculated by Monte Carlo simulation. Our own calculations resulted in critical values at the 0.05 level of 0.46 for $T_1$ and 2.50 for $T_2$.

## 11.3 Contiguous Alternatives

For each $n \geq 1$, let $X_{n1}, \ldots, X_{nn}$ be i.i.d. random elements in a measurable space $(\mathcal{X}, A)$. Let $P$ denote the common probability distribution under the "null hypothesis," and let $P_n$ be a "contiguous alternative hypothesis" distribution satisfying

$$\int \left[\sqrt{n}(dP_n^{1/2} - dP^{1/2}) - \frac{1}{2}h dP^{1/2}\right]^2 \to 0, \tag{11.4}$$

as $n \to \infty$, for some measurable function $h : \mathcal{X} \mapsto \mathbb{R}$. The following lemma, which is part of Lemma 3.10.11 of VW and which we give without proof, provides some properties for $h$:

LEMMA 11.11 *If the sequence of probability measures $P_n$ satisfy (11.4), then necessarily $Ph = 0$ and $Ph^2 < \infty$.*

The following theorem gives very general weak convergence properties of the empirical process under the contiguous alternative $P_n$. Such weak convergence will be useful for studying efficiency of tests in Chapter 18. This is Theorem 3.10.12 of VW which we give without proof:

THEOREM 11.12 *Let $\mathcal{F}$ be a $P$-Donsker class of measurable functions with $\|P\|_{\mathcal{F}} < \infty$, and assume the sequence of probability measures $P_n$ satisfies (11.4). Then $\sqrt{n}(\mathbb{P}_n - P)$ 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. If, moreover, $\|P_n f^2\|_{\mathcal{F}} = O(1)$, then $\|\sqrt{n}(P_n - P)f - Pfh\|_{\mathcal{F}} \to 0$ and $\sqrt{n}(\mathbb{P}_n - P_n)$ converges under $P_n$ in distribution to $\mathbb{G}$.*

We now present a bootstrap result for contiguous alternatives. For i.i.d. nonnegative random weights $\xi_1, \ldots, \xi_n$ with mean $0 < \mu < \infty$ and variance $0 < \tau^2 < \infty$, recall the bootstrapped empirical measures $\tilde{\mathbb{P}}_n f = n^{-1} \sum_{i=1}^n (\xi_i/\bar{\xi}) f(X_i)$ and $\tilde{\mathbb{G}}_n = \sqrt{n}(\mu/\tau)(\tilde{\mathbb{P}}_n - \mathbb{P}_n)$ from Sections 2.2.3 and 10.1. We define several new symbols: $\overset{P_n}{\to}$ and $\overset{P_n}{\rightsquigarrow}$ denote convergence in probability and weak convergence, respectively, under the distribution $P_n^n$. In addition, $\hat{X}_n \underset{M}{\overset{P_n}{\rightsquigarrow}} X$ in a metric space $\mathbb{D}$ denotes conditional bootstrap convergence in probability under $P_n$, i.e., $\sup_{g \in BL_1} |\mathrm{E}_M g(\hat{X}_n) - \mathrm{E} g(X)| \overset{P_n}{\to} 0$ and $\mathrm{E}_M g(\hat{X}_n)^* - \mathrm{E}_M g(\hat{X}_n)_* \overset{P_n}{\to} 0$, for all $g \in BL_1$, where $BL_1$ is the same bounded Lipschitz function space defined in Section 2.2.3 and the subscript $M$ denotes taking the expectation over $\xi_1, \ldots, \xi_n$ conditional on the data. Note that we require the weights $\xi_1, \ldots, \xi_n$ to have the same distribution and independence from $X_{n1}, \ldots, X_{nn}$ under both $P_n$ and $P$.

THEOREM 11.13 *Let $\mathcal{F}$ be a $P$-Donsker class of measurable functions, let $P_n$ satisfy (11.4), and assume*

$$\lim_{M \to \infty} \limsup_{n \to \infty} P_n (f - Pf)^2 1\{|f - Pf| > M\} \quad = \quad 0 \tag{11.5}$$

*for all $f \in \mathcal{F}$. Also let $\xi_1, \ldots, \xi_n$ be i.i.d. nonnegative random variables, independent of $X_{n1}, \ldots, X_{nn}$, with mean $0 < \mu < \infty$, variance $0 < \tau^2 < \infty$, and with $\|\xi_1\|_{2,1} < \infty$. Then $\tilde{\mathbb{G}}_n \underset{\xi}{\overset{P_n}{\rightsquigarrow}} \mathbb{G}$ in $\ell^\infty(\mathcal{F})$ and $\tilde{\mathbb{G}}_n$ is asymptotically measurable.*

**Proof.** Let $\eta_i \equiv \tau^{-1}(\xi_i - \mu)$, $i = 1, \ldots, n$, and note that

$$
\begin{aligned}
(11.6) \qquad \tilde{\mathbb{G}}_n &= n^{-1/2}(\mu/\tau)\sum_{i=1}^{n}(\xi_i/\bar{\xi}-1)\delta_{X_i} \\
&= n^{-1/2}(\mu/\tau)\sum_{i=1}^{n}(\xi_i/\bar{\xi}-1)(\delta_{X_i}-P) \\
&= n^{-1/2}\sum_{i=1}^{n}\eta_i(\delta_{X_i}-P) \\
&\quad + \left(\frac{\mu}{\bar{\xi}}-1\right)n^{-1/2}\sum_{i=1}^{n}\eta_i(\delta_{X_i}-P) \\
&\quad + \left(\frac{\mu}{\tau}\right)\left(\frac{\mu}{\bar{\xi}}-1\right)n^{-1/2}\sum_{i=1}^{n}(\delta_{X_i}-P).
\end{aligned}
$$

Since $\mathcal{F}$ is $P$-Donsker, we also have that $\dot{\mathcal{F}} \equiv \{f - Pf : f \in \mathcal{F}\}$ is $P$-Donsker. Thus by the unconditional multiplier central limit theorem (Theorem 10.1), we have that $\eta \cdot \mathcal{F}$ is also $P$-Donsker. Now, by the fact that $\|P(f-Pf)\|_{\mathcal{F}} = 0$ (trivially) combined with Theorem 11.12, both $n^{-1/2}\sum_{i=1}^{n}\eta_i(\delta_{X_i}-P) \overset{P_n}{\rightsquigarrow} \mathbb{G}(f)$ and $n^{-1/2}\sum_{i=1}^{n}(\delta_{X_i}-P) \overset{P_n}{\rightsquigarrow} \mathbb{G}(f)+P(f-Pf)h$ in $\ell^\infty(\mathcal{F})$. The reason the first limiting process has mean zero is because $\eta$ is independent of $X$ and thus $P\eta(f-Pf)h = 0$ for all $f \in \mathcal{F}$. Thus the last two terms in (11.6) $\overset{P_n}{\to} 0$ and $\tilde{\mathbb{G}}_n \overset{P_n}{\rightsquigarrow} \mathbb{G}$ in $\ell^\infty(\mathcal{F})$. This now implies the unconditional asymptotic tightness and desired asymptotic measurability of $\tilde{\mathbb{G}}_n$.

By the same kinds of arguments we used in the proof of Theorem 10.4, all we need to verify now is that all finite dimensional collections $f_1, \ldots, f_m \in \mathcal{F}$ converge under $P_n$ in distribution, conditional on the data, to the appropriate limiting Gaussian process. Accordingly, let $Z_i = (f_1(X_i) - Pf_1, \ldots, f_m(X_i) - Pf_m)'$, $i = 1, \ldots, n$. What we need to show is that $n^{-1/2}\sum_{i=1}^{n}\eta_i Z_i$ converges weakly under $P_n$, conditional on the $Z_1, \ldots, Z_n$, to a mean zero Gaussian process with variance $\Sigma \equiv PZ_1Z_1'$. By Lemma 10.5, we are done if we can verify that $n^{-1}\sum_{i=1}^{n}Z_iZ_i' \overset{P_n}{\to} \Sigma$ and $\max_{1\le i\le n}\|Z_i\|/\sqrt{n} \overset{P_n}{\to} 0$.

By the assumptions of the theorem,

$$
\limsup_{n\to\infty}\{E_n(M) \equiv P_n[Z_1Z_1'1\{\|Z_1\| > M\}]\} \to 0
$$

as $M \to \infty$. Note also that (11.4) can be shown to imply that $P_nh \to Ph$, as $n \to \infty$, for any bounded $h$ (this verification is saved as an exercise). Thus, for any $M < \infty$, $P_nZ_1Z_1'1\{\|Z_1\| \le M\}$ convergences to $PZ_1Z_1'1\{\|Z_1\| \le M\}$. Since $M$ is arbitrary, this convergence continues to hold if $M$ is replaced by a sequence $M_n$ going to infinity slowly enough. Accordingly,

$$\begin{aligned} n^{-1}\sum_{i=1}^{n} Z_i Z_i' &= n^{-1}\sum_{i=1}^{n} Z_i Z_i' 1\{\|Z_i\| > M_n\} \\ &\quad + n^{-1}\sum_{i=1} Z_i Z_i' 1\{\|Z_i\| \le M_n\} \\ &\xrightarrow{P_n} PZ_1 Z_1', \end{aligned}$$

as $n \to \infty$. Now we also have

$$\begin{aligned} \max_{1\le i\le n} \frac{\|Z_i\|}{\sqrt{n}} &= \sqrt{\max_{1\le i\le n} \frac{\|Z_i\|^2}{n}} \\ &\le \sqrt{\max_{1\le i\le n} \frac{\|Z_i\|^2}{n} 1\{\|Z_i\| \le M\} + \frac{1}{n}\sum_{i=1}^{n} \|Z_i\|^2 1\{\|Z_i\| > M\}}. \end{aligned}$$

The first term under the last square root sign $\xrightarrow{P_n} 0$ trivially, while the expectation under $P_n$ of the second term, $E_n(M)$, goes to zero as $n \to \infty$ and $M \to \infty$ sufficiently slowly with $n$, as argued previously. Thus $\max_{1\le i\le n} \|Z_i\|/\sqrt{n} \xrightarrow{P_n} 0$, and the proof is complete.□

We close this section with the following additional useful consequences of contiguity:

THEOREM 11.14 *Let $Y_n = Y_n(X_{n1}, \ldots, X_{nn})$ and assume that (11.4) holds. Then the following are true:*

*(i) If $Y_n \xrightarrow{\mathrm{P}} 0$ under the sequence of distributions $P^n$, then $Y_n \xrightarrow{P_n} 0$.*

*(ii) If $Y_n$ is asymptotically tight under the sequence of distributions $P^n$, then $Y_n$ is also asymptotically tight under $P_n^n$.*

**Proof.** In addition to the result of Lemma 11.11, Lemma 3.10.11 of VW yields that, under (11.4),

$$\sum_{i=1}^{n} \log \frac{dP_n}{dP}(X_{ni}) = G_n - \frac{1}{2}Ph^2 + R_n,$$

where $G_n \equiv \sqrt{n}\mathbb{P}_n h(X)$ and $R_n$ goes to zero under both $P^n$ and $P_n^n$. Note that by Theorem 11.12, $G_n \xrightarrow{P_n} \mathbb{G}h + Ph^2$.

We first prove (i). Note that for every $\delta > 0$, there exists an $M < \infty$ such that $\limsup_{n\to\infty} P_n^n(|G_n| > M) \le \delta$. For any $\epsilon > 0$, let $g_n(\epsilon) \equiv 1\{\|Y_n\|^* > \epsilon\}$, and note that $g_n(\epsilon)$ is measurable under both $P^n$ and $P_n^n$. Thus, for any $\epsilon > 0$,

$$\begin{aligned} P_n^n g_n(\epsilon) &\le P_n^n\left[g_n(\epsilon) 1\{|G_n| \le M, |R_n| \le \epsilon\}\right] + P_n^n(|G_n| > M) \\ &\quad + P_n^n(|R_n| > \epsilon) \\ &\le P^n\left[g_n(\epsilon) e^{M - Ph^2/2 - \epsilon}\right] + P_n^n(|G_n| > M) + P_n^n(|R_n| > \epsilon) \\ &\to 0 + \delta + 0, \end{aligned}$$

as $n \to \infty$. Thus (i) follows since both $\delta$ and $\epsilon$ were arbitrary.

We now prove (ii). Fix $\eta > 0$, and note that by recycling the above arguments, $H_n \equiv \prod_{i=1}^{n}(dP_n/dP)(X_{ni})$ is asymptotically bounded in probability under both $P^n$ and $P_n^n$. Thus there exists an $H < \infty$ such that $\limsup_{n\to\infty} \mathrm{P}(H_n > H) \leq \eta/2$. Since $Y_n$ is asymptotically tight under $P^n$, there exists a compact set $K$ such that $\limsup_{n\to\infty} \mathrm{P}(Y_n \in (\mathcal{X} - K^\delta)^*) \leq \eta/(2H)$ under $P^n$, for every $\delta > 0$, where $K^\delta$ is the $\delta$-enlargement of $K$ as defined in Section 7.2.1 and superscript $*$ denotes a set that is the minimum measurable covering under both $P^n$ and $P_n^n$. Hence, under $P_n^n$,

$$\begin{aligned}
\limsup_{n\to\infty} \mathrm{P}(Y_n \in (\mathcal{X} - K^\delta)^*) &= \limsup_{n\to\infty} \int 1\{Y_n \in (\mathcal{X} - K^\delta)^*\} dP_n^n \\
&\leq H \limsup_{n\to\infty} \int 1\{Y_n \in (\mathcal{X} - K^\delta)^*\} dP^n \\
&\quad + \limsup_{n\to\infty} \mathrm{P}(H_n > H) \\
&\leq H\frac{\eta}{2H} + \frac{\eta}{2} \\
&= \eta,
\end{aligned}$$

for every $\delta > 0$. Thus $Y_n$ is asymptotically tight under $P_n^n$ since $\eta$ was arbitrary. Hence (ii) follows.□

## 11.4 Sums of Independent but not Identically Distributed Stochastic Processes

In this section, we are interested in deriving the limiting distribution of sums of the form $\sum_{i=1}^{m_n} f_{ni}(\omega, t)$, where the real-valued stochastic processes $\{f_{ni}(\omega, t), t \in T, 1 \leq i \leq m_n\}$, for all integers $n \geq 1$, are independent within rows on the probability space $(\Omega, \mathcal{A}, P)$ for some index set $T$. In addition to a central limit theorem, we will present a multiplier bootstrap result to aid in inference. An example using these techniques will be presented in the upcoming case studies II chapter. Throughout this section, function arguments or subscripts will sometimes be suppressed for notational clarity.

### *11.4.1 Central Limit Theorems*

The notation and set-up are similar to that found in Pollard (1990) and Kosorok (2003). A slightly more general approach to the same question can be found in Chapter 2.11 of VW. Part of the generality in VW is the ability to utilize bracketing entropy in addition to uniform entropy for establishing tightness. An advantage of Pollard's approach, on the other hand, is that total boundedness of the index set $T$ is a conclusion rather than a condition.

Both approaches have their merits and appear to be roughly equally useful in practice.

We need to introduce a few measurability conditions that are different from but related to conditions introduced in previous chapters. The first condition is *almost measurable Suslin*: Call a triangular array $\{f_{ni}(\omega,t), t \in T\}$ almost measurable Suslin (AMS) if for all integers $n \geq 1$, there exists a Suslin topological space $T_n \subset T$ with Borel sets $\mathcal{B}_n$ such that

(i)
$$\mathrm{P}^* \left( \sup_{t\in T} \inf_{s\in T_n} \sum_{i=1}^{m_n} \left(f_{ni}(\omega,s) - f_{ni}(\omega,t)\right)^2 > 0 \right) = 0,$$

(ii) For $i = 1 \ldots m_n$, $f_{ni} : \Omega \times T_n \mapsto \mathbb{R}$ is $\mathcal{A} \times \mathcal{B}_n$-measurable.

The second condition is stronger yet seems to be more easily verified in applications: Call a triangular array of processes $\{f_{ni}(\omega,t), t \in T\}$ *separable* if for every integer $n \geq 1$, there exists a countable subset $T_n \subset T$ such that

$$\mathrm{P}^* \left( \sup_{t\in T} \inf_{s\in T_n} \sum_{i=1}^{m_n} \left(f_{ni}(\omega,s) - f_{ni}(\omega,t)\right)^2 > 0 \right) = 0.$$

The following lemma shows that separability implies AMS:

LEMMA 11.15 *If the triangular array of stochastic processes* $\{f_{ni}(\omega,t), t \in T\}$ *is separable, then it is AMS.*

**Proof.** The discrete topology applied to $T_n$ makes it into a Suslin topology by countability, with resulting Borel sets $\mathcal{B}_n$. For $i = 1 \ldots m_n$, $f_{ni} : \Omega \times T_n \mapsto \mathbb{R}$ is $\mathcal{A} \times \mathcal{B}_n$-measurable since, for every $\alpha \in \mathbb{R}$,

$$\{(\omega,t) \in \Omega \times T_n : f_{ni}(\omega,t) > \alpha\} = \bigcup_{s\in T_n} \{(\omega,s) : f_{ni}(\omega,s) > \alpha\},$$

and the right-hand-side is a countable union of $\mathcal{A} \times \mathcal{B}_n$-measurable sets.□

The forgoing measurable Suslin condition is closely related to the definition given in Example 2.3.5 of VW, while the definition of separable arrays is similar in spirit to the definition of separable stochastic processes given in the discussion preceding Lemma 7.2 in Section 7.1 above. The modifications of these definitions presented in this section have been made to accommodate nonidentically distributed arrays for a broad scope of statistical applications. However, finding the best possible measurability conditions was not the primary goal.

We need the following definition of *manageability* (Definition 7.9 of Pollard, 1990, with minor modification). First, for any set $A \in \mathbb{R}^m$, let $D_m(x, A)$ be the packing number for the set $A$ at Euclidean distance $x$, i.e., the largest $k$ such that there exist $k$ points in $A$ with the smallest Euclidean distance between any two distinct points being greater than $x$. Also let $\mathcal{F}_{n\omega} \equiv$

$\{[f_{n1}(\omega,t),\ldots,f_{nm_n}(\omega,t)] \in \mathbb{R}^{m_n} : t \in T\}$; and for any vectors $u, v \in \mathbb{R}^m$, $u \odot v \in \mathbb{R}^m$ is the pointwise product and $\|\cdot\|$ denotes Euclidean distance. A triangular array of processes $\{f_{ni}(\omega,t)\}$ is manageable, with respect to the envelopes $F_n(\omega) \equiv [F_{n1}(\omega),\ldots,F_{nm_n}(\omega)] \in \mathbb{R}^{m_n}$, if there exists a deterministic function $\lambda$ (the *capacity bound*) for which

(i) $\int_0^1 \sqrt{\log \lambda(x)}dx < \infty$,

(ii) there exists $N \subset \Omega$ such that $\mathrm{P}^*(N) = 0$ and for each $\omega \notin N$,

$$D_{m_n}\left(x \left\|\alpha \odot F_n(\omega)\right\|, \alpha \odot \mathcal{F}_{n\omega}\right) \leq \lambda(x),$$

for $0 < x \leq 1$, all vectors $\alpha \in \mathbb{R}^{m_n}$ of nonnegative weights, all $n \geq 1$, and where $\lambda$ does not depend on $\omega$ or $n$.

We now state a minor modification of Pollard's Functional Central Limit Theorem for the stochastic process sum

$$X_n(\omega,t) \equiv \sum_{i=1}^{m_n} \left[f_{ni}(\omega,t) - \mathrm{E}f_{ni}(\cdot,t)\right].$$

The modification is the inclusion of a sufficient measurability requirement which was omitted in Pollard's (1990) version of the theorem.

THEOREM 11.16 *Suppose the triangular array* $\{f_{ni}(\omega,t),\ i = 1,\ldots,m_n,\ t \in T\}$ *consists of independent processes within rows, is AMS, and satisfies:*

*(A)* *the* $\{f_{ni}\}$ *are manageable, with envelopes* $\{F_{ni}\}$ *which are also independent within rows;*

*(B)* $H(s,t) = \lim_{n\to\infty} \mathrm{E}X_n(s)X_n(t)$ *exists for every* $s,t \in T$*;*

*(C)* $\limsup_{n\to\infty} \sum_{i=1}^{m_n} \mathrm{E}^* F_{ni}^2 < \infty$*;*

*(D)* $\lim_{n\to\infty} \sum_{i=1}^{m_n} \mathrm{E}^* F_{ni}^2 1\{F_{ni} > \epsilon\} = 0$*, for each* $\epsilon > 0$*;*

*(E)* $\rho(s,t) = \lim_{n\to\infty} \rho_n(s,t)$*, where*

$$\rho_n(s,t) \equiv \left(\sum_{i=1}^{m_n} \mathrm{E}\left|f_{ni}(\cdot,s) - f_{ni}(\cdot,t)\right|^2\right)^{1/2},$$

*exists for every* $s,t \in T$*, and for all deterministic sequences* $\{s_n\}$ *and* $\{t_n\}$ *in* $T$*, if* $\rho(s_n,t_n) \to 0$ *then* $\rho_n(s_n,t_n) \to 0$*.*

*Then*

*(i)* $T$ *is totally bounded under the* $\rho$ *pseudometric;*

*(ii)* $X_n$ *converges weakly on* $\ell^\infty(T)$ *to a tight mean zero Gaussian process* $X$ *concentrated on* $UC(T,\rho)$*, with covariance* $H(s,t)$*.*

The proof is given in Kosorok (2003), who relies on Chapter 10 of Pollard (1990) for some of the steps. We omit the details. We will use this theorem when discussing function classes changing with $n$ later in this chapter, as well as in one of the case studies of Chapter 15. One can think of this theorem as a Lindeberg central limit theorem for stochastic processes, where Condition (D) is a modified Lindeberg condition.

Note that the manageability Condition (A) is an entropy condition quite similar to the BUEI condition of Chapter 9. Pollard (1990) discussed several methods and preservation results for establishing manageability, including *bounded pseudodimension classes* which are very close in spirit to VC-classes of functions (see Chapter 4 of Pollard): The set $\mathcal{F}_n \subset \mathbb{R}^n$ has pseudodimension of at most $V$ if, for every point $t \in \mathbb{R}^{V+1}$, no *proper coordinate projection* of $\mathcal{F}_n$ can surround $t$. A proper coordinate projection $\mathcal{F}_n^k$ is obtained by choosing a subset $i_1, \ldots, i_k$ of indices $1, \ldots, m_n$, where $k \leq m_n$, and then letting $\mathcal{F}_n^k = \left\{ f_{ni_1}(t_1), \ldots, f_{ni_k}(t_k) : (t_1, \ldots, t_k) \in \mathbb{R}^k \right\}$.

It is not hard to verify that if $f_{n1}(t), \ldots, f_{nm_n}(t)$ are always monotone increasing functions in $t$, then the resulting $\mathcal{F}_n$ has pseudodimension 1. This happens because all two-dimensional projections always form monotone increasing trajectories, and thus can never surround any point in $\mathbb{R}^2$. The proof is almost identical to the proof of Lemma 9.10. By Theorem 4.8 of Pollard (1990), every triangular array of stochastic processes for which $\mathcal{F}_n$ has pseudodimension bounded above by $V < \infty$, for all $n \geq 1$, is manageable. As verified in the following theorem, complicated manageable classes can be built up from simpler manageable classes:

THEOREM 11.17 *Let $f_{n1}, \ldots, f_{nm_n}$ and $g_{n1}, \ldots, g_{nm_n}$ be manageable arrays with respective index sets $T$ and $U$ and with respective envelopes $F_{n1}, \ldots, F_{nm_n}$ and $G_{n1}, \ldots, G_{nm_n}$. Then the following are true:*

*(i)* $\{f_{n1}(t) + g_{n1}(u), \ldots, f_{nm_n}(t) + g_{nm_n}(u) : (t,u) \in T \times U\}$, *is manageable with envelopes* $F_{n1} + G_{n1}, \ldots, F_{nm_n} + G_{nm_n}$;

*(ii)* $\{f_{n1}(t) \wedge g_{n1}(u), \ldots, f_{nm_n}(t) \wedge g_{nm_n}(u) : (t,u) \in T \times U\}$ *is manageable with envelopes* $F_{n1} + G_{n1}, \ldots, F_{nm_n} + G_{nm_n}$;

*(iii)* $\{f_{n1}(t) \vee g_{n1}(u), \ldots, f_{nm_n}(t) \vee g_{nm_n}(u) : (t,u) \in T \times U\}$ *is manageable with envelopes* $F_{n1} + G_{n1}, \ldots, F_{nm_n} + G_{nm_n}$;

*(iv)* $\{f_{n1}(t) g_{n1}(u), \ldots, f_{nm_n}(t) g_{nm_n}(u) : (t,u) \in T \times U\}$ *is manageable with envelopes* $F_{n1} G_{n1}, \ldots, F_{nm_n} G_{nm_n}$.

**Proof.** For any vectors $x_1, x_2, y_1, y_2 \in \mathbb{R}^{m_n}$, it is easy to verify that $\|x_1 \square y_1 - x_2 \square y_2\| \leq \|x_1 - x_2\| + \|y_1 - y_2\|$, where $\square$ is any one of the operations $\wedge$, $\vee$, or $+$. Thus

$$\begin{aligned} N_{m_n}\left(\epsilon \|\alpha \odot (F_n + G_n)\|, \alpha \odot \mathcal{F}_n \square \mathcal{G}_n\right) \quad \leq \quad & N_{m_n}\left(\epsilon \|\alpha \odot F_n\|, \alpha \odot \mathcal{F}_n\right) \\ & \times N_{m_n}\left(\epsilon \|\alpha \odot G_n\|, \alpha \odot \mathcal{G}_n\right), \end{aligned}$$

for any $0 < \epsilon \leq 1$ and all vectors $\alpha \in \mathbb{R}^{m_n}$ of nonnegative weights, where $N_{m_n}$ denotes the covering number version of $D_{m_n}$, $\mathcal{G}_n \equiv \{g_{n1}(u), \ldots, g_{nm_n}(u) : u \in U\}$, and $\mathcal{F}_n \Box \mathcal{G}_n$ has the obvious interpretation. Thus Parts (i)–(iii) of the theorem follow by the relationship between packing and covering numbers discussed in Section 8.1.2 in the paragraphs preceding Theorem 8.4.

Proving Part (iv) requires a slightly different approach. Let $x_1, x_2, y_1, y_2$ by any vectors in $\mathbb{R}^{m_n}$, and let $\tilde{x}, \tilde{y} \in \mathbb{R}^{m_n}$ be any nonnegative vectors such that both $\tilde{x} - [x_1 \vee x_2 \vee (-x_1) \vee (-x_2)]$ and $\tilde{y} - [y_1 \vee y_2 \vee (-y_1) \vee (-y_2)]$ are nonnegative. It is not hard to verify that $\|x_1 \odot y_1 - x_2 \odot y_2\| \leq \|\tilde{y} \odot x_1 - \tilde{y} \odot x_2\| + \|\tilde{x} \odot y_1 - \tilde{x} \odot y_2\|$. From this, we can deduce that

$$\begin{aligned} & N_{m_n}\left(2\epsilon\|\alpha \odot F_n \odot G_n\|, \alpha \odot \mathcal{F}_n \odot \mathcal{G}_n\right) \\ & \quad \leq \quad N_{m_n}\left(\epsilon\|\alpha \odot F_n \odot G_n\|, \alpha \odot G_n \odot \mathcal{F}_n\right) \\ & \qquad \times N_{m_n}\left(\epsilon\|\alpha \odot F_n \odot G_n\|, \alpha \odot F_n \odot \mathcal{G}_n\right) \\ & \quad = \quad N_{m_n}\left(\epsilon\|\alpha' \odot F_n\|, \alpha' \odot \mathcal{F}_n\right) \times N_{m_n}\left(\epsilon\|\alpha'' \odot G_n\|, \alpha'' \odot \mathcal{G}_n\right), \end{aligned}$$

for any $0 < \epsilon \leq 1$ and all vectors $\alpha \in \mathbb{R}^{m_n}$ of nonnegative weights, where $\alpha' \equiv \alpha \odot G_n$ and $\alpha'' \equiv \alpha \odot F_n$. Since capacity bounds do not depend on the nonnegative weight vector (either $\alpha$, $\alpha'$ or $\alpha''$), Part (iv) now follows, and the theorem is proved.□

### *11.4.2 Bootstrap Results*

We now present a weighted bootstrap for inference about the limiting process $X$ of Theorem 11.16. The basic idea shares some similarities with the wild bootstrap (Praestgaard and Wellner, 1993). Let $Z \equiv \{z_i, i \geq 1\}$ be a sequence of random variables satisfying

(F) The $\{z_i\}$ are independent and identically distributed, on the probability space $\{\Omega_z, \mathcal{A}_z, \Pi_z\}$, with mean zero and variance 1.

Denote $\mu_{ni}(t) \equiv \mathrm{E} f_{ni}(\cdot, t)$, and let $\hat{\mu}_{ni}(t)$ be estimators of $\mu_{ni}(t)$. The weighted bootstrapped process we propose for inference is

$$\hat{X}_{n\omega}(t) \equiv \sum_{i=1}^{m_n} z_i \left[f_{ni}(\omega, t) - \hat{\mu}_{ni}(\omega, t)\right],$$

which is defined on the product probability space $\{\Omega, \mathcal{A}, \Pi\} \times \{\Omega_z, \mathcal{A}_z, \Pi_z\}$, similar to what was done in Chapter 9 for the weighted bootstrap in the i.i.d. case. What is unusual about this bootstrap is the need to estimate the $\mu_{ni}$ terms: this need is a consequence of the terms being non-identically distributed.

The proposed method of inference is to resample $\hat{X}_n$, using many realizations of $z_1, \ldots, z_{m_n}$, to approximate the distribution of $X_n$. The following theorem gives us conditions under which this procedure is asymptotically valid:

THEOREM 11.18 *Suppose the triangular array* $\{f_{ni}\}$ *satisfies the conditions of Theorem 11.16 and the sequence* $\{z_i, i \geq 1\}$ *satisfies Condition (F) above. Suppose also that the array of estimators* $\{\hat{\mu}_{ni}(\omega, t), t \in T, 1 \leq i \leq m_n, n \geq 1\}$ *is AMS and satisfies the following:*

*(G)* $\sup_{t \in T} \sum_{i=1}^{m_n} \left[\hat{\mu}_{ni}(\omega, t) - \mu_{ni}(t)\right]^2 = o_P(1)$;

*(H) the stochastic processes* $\{\hat{\mu}_{ni}(\omega, t)\}$ *are manageable with envelopes* $\left\{\hat{F}_{ni}(\omega)\right\}$;

*(I)* $k \vee \sum_{i=1}^{m_n} \left[\hat{F}_{ni}(\omega)\right]^2$ *converges to* $k$ *in outer probability as* $n \to \infty$, *for some* $k < \infty$.

*Then the conclusions of Theorem 11.16 obtain,* $\hat{X}_n$ *is asymptotically measurable, and* $\hat{X}_n \underset{Z}{\overset{\mathrm{P}}{\rightsquigarrow}} X$.

The main idea of the proof is to first study the conditional limiting distribution of $\tilde{X}_{n\omega}(t) \equiv \sum_{i=1}^{m_n} z_i \left[f_{ni}(\omega, t) - \mu_{ni}(t)\right]$, and then show that the limiting result is unchanged after replacing $\mu_{ni}$ with $\hat{\mu}_{ni}$. The first step is summarized in the following theorem:

THEOREM 11.19 *Suppose the triangular array* $\{f_{ni}\}$ *satisfies the conditions of Theorem 11.16 and the sequence* $\{z_i, i \geq 1\}$ *satisfies Condition (F) above. Then the conclusions of Theorem 11.16 obtain,* $\tilde{X}_n$ *is asymptotically measurable, and* $\tilde{X}_n \underset{Z}{\overset{\mathrm{P}}{\rightsquigarrow}} X$.

An interesting step in the proof of this theorem is verifying that manageability of the triangular array $z_1 f_{n1}, \ldots, z_{m_n} f_{nm_n}$ follows directly from manageability of $f_{n1}, \ldots, f_{nm_n}$. We now demonstrate this. For vectors $u \in \mathbb{R}^{m_n}$, let $|u|$ denote pointwise absolute value and $\mathrm{sign}(u)$ denote pointwise sign. Now, for any nonnegative $\alpha \in \mathbb{R}^{m_n}$,

$$\begin{aligned} & D_{m_n}\left(x \left\|\alpha \odot |z_n| \odot F_n(\omega)\right\|, \alpha \odot z_n \odot \mathcal{F}_{n\omega}\right) \\ &= D_{m_n}\left(x \left\|\tilde{\alpha} \odot F_n(\omega)\right\|, \tilde{\alpha} \odot \mathrm{sign}(z_n) \odot \mathcal{F}_{n\omega}\right) \\ &= D_{m_n}\left(x \left\|\tilde{\alpha} \odot F_n(\omega)\right\|, \tilde{\alpha} \odot \mathcal{F}_{n\omega}\right), \end{aligned}$$

where $z_n \equiv \{z_1, \ldots, z_{m_n}\}^T$, since the absolute value of the $\{z_i\}$ can be absorbed into the $\alpha$ to make $\tilde{\alpha}$ and since any coordinate change of sign does not effect the geometry of $\mathcal{F}_{n\omega}$. Thus the foregoing triangular array is manageable with envelopes $\{|z_i| F_{ni}(\omega)\}$. The remaining details of the proofs of both Theorems 11.18 and 11.19, which we omit here, can be found in Kosorok (2003).

## 11.5 Function Classes Changing with $n$

We now return to the i.i.d. empirical process setting where the i.i.d. observations $X_1, X_2, \ldots$ are drawn from a measurable space $\{\mathcal{X}, \mathcal{A}\}$, with probability measure $P$. What is new, however, is that we allow the function class to depend on $n$. Specifically, we assume the function class has the form $\mathcal{F}_n \equiv \{f_{n,t} : t \in T\}$, where the functions $x \mapsto f_{n,t}(x)$ are indexed by a fixed $T$ but are allowed to change with sample size $n$. Note that this trivially includes the standard empirical process set-up with an arbitrary but fixed function class $\mathcal{F}$ by setting $T = \mathcal{F}$ and $f_{n,t} = t$ for all $n \geq 1$ and $t \in \mathcal{F}$. The approach we take is to specialize the results of Section 11.4 after replacing manageability with a bounded uniform entropy integral condition. An alternative approach which can utilize either uniform or bracketing entropy is given in Section 2.11.3 of VW, but we do not pursue this second approach here.

Let $X_n(t) \equiv n^{-1/2} \sum_{i=1}^n (f_{n,t}(X_i) - Pf_{n,t})$, for all $t \in T$, and let $F_n$ be an envelope for $\mathcal{F}_n$. We say that the sequence $\mathcal{F}_n$ is AMS if for all $n \geq 1$, there exists a Suslin topological space $T_n \subset T$ with Borel sets $\mathcal{B}_n$ such that

$$\text{(11.7)} \qquad \mathrm{P}^* \left( \sup_{t \in T} \inf_{s \in T_n} |f_{n,s}(X_1) - f_{n,t}(X_1)| > 0 \right) \quad = \quad 0$$

and $f_{n,\cdot} : \mathcal{X} \times T_n \mapsto \mathbb{R}$ is $\mathcal{A} \times \mathcal{B}_n$-measurable. Moreover, the sequence $\mathcal{F}_n$ is said to be separable if, for all $n \geq 1$, there exists a countable subset $T_n \subset T$ such that (11.7) holds. The arguments in the proof of Lemma 11.15 verify that separability implies AMS as is true for the more general setting of Section 11.4.

We also require the following bounded uniform entropy integral condition:

$$\text{(11.8)} \qquad \limsup_{n \to \infty} \sup_Q \int_0^1 \sqrt{\log N \left(\epsilon \|F_n\|_{Q,2}, \mathcal{F}_n, L_2(Q)\right)} d\epsilon \quad < \quad \infty,$$

where, for each $n \geq 1$, $\sup_Q$ is the supremum taken over all finitely discrete probability measures $Q$ with $\|F_n\|_{Q,2} > 0$. We are now ready to present the following functional central limit theorem:

THEOREM 11.20 *Suppose $\mathcal{F}_n$ is AMS and the following hold:*

*(A) $\mathcal{F}_n$ satisfies (11.8) with envelop $F_n$;*

*(B) $H(s,t) = \lim_{n \to \infty} \mathrm{E} X_n(s) X_n(t)$ for every $s, t \in T$;*

*(C) $\limsup_{n \to \infty} \mathrm{E}^* F_n^2 < \infty$;*

*(D) $\lim_{n \to \infty} \mathrm{E}^* F_n^2 1\{F_n > \epsilon \sqrt{n}\} = 0$, for each $\epsilon > 0$;*

*(E) $\rho(s,t) = \lim_{n \to \infty} \rho_n(s,t)$, where $\rho_n(s,t) \equiv \sqrt{\mathrm{E}|f_{n,s}(X_1) - f_{n,t}(X_2)|}$, exists for every $s, t \in T$, and for all deterministic sequences $\{s_n\}$ and $\{t_n\}$ in $T$, if $\rho(s_n, t_n) \to 0$ then $\rho_n(s_n, t_n) \to 0$.*

*Then*

*(i) $T$ is totally bounded under the $\rho$ pseudometric;*

*(ii) $X_n$ converges weakly in $\ell^\infty(T)$ to a tight, mean zero Gaussian process $X$ concentrated on $UC(T,\rho)$, with covariance $H(s,t)$.*

**Proof.** The proof consists in showing that the current setting is just a special case of Theorem 11.16. Specifically, we let $f_{ni}(t) = f_{n,t}(X_i)$ and $m_n = n$, and we study the array $\{f_{ni}(t), t \in T\}$. First, it is clear that $\mathcal{F}_n$ being AMS implies that $\{f_{ni}(t), t \in T\}$ is AMS. Now let

$$\tilde{\mathcal{F}}_n \equiv \{[f_{n1}(t), \ldots, f_{nn}(t)] \in \mathbb{R}^n : t \in T\}$$

and $\tilde{F}_n \equiv \left[\tilde{F}_{n1}, \ldots, \tilde{F}_{nn}\right]$, where $\tilde{F}_{ni} \equiv F_n(X_i)/\sqrt{n}$; and note that for any $\alpha \in \mathbb{R}^n$,

$$D_n\left(\epsilon\|\alpha \odot \tilde{F}_n\|, \alpha \odot \tilde{\mathcal{F}}_n\right) \leq D\left(\epsilon\|F_n\|_{Q_\alpha,2}, \mathcal{F}_n, L_2(Q_\alpha)\right),$$

where $Q_\alpha \equiv (n\|\alpha\|)^{-1}\sum_{i=1}^n \alpha_i^2 \delta_{X_i}$ is a finitely discrete probability measure. Thus, by the relationship between packing and covering numbers given in Section 8.1.2, we have that if we let

$$\lambda(x) = \limsup_{n\to\infty} \sup_Q N\left(x\|F_n\|_{Q,2}/2, \mathcal{F}_n, L_2(Q)\right),$$

where $\sup_Q$ is taken over all finitely discrete probability measures, then Condition 11.8 implies that

$$D_n\left(\epsilon\|\alpha \odot \tilde{F}_n\|, \alpha \odot \tilde{\mathcal{F}}_n\right) \leq \lambda(\epsilon),$$

for all $0 < \epsilon \leq 1$, all vectors $\alpha \in \mathbb{R}^n$ of nonnegative weights, and all $n \geq 1$; and that $\int_0^1 \sqrt{\log \lambda(\epsilon)}d\epsilon < \infty$. Note that without loss of generality, we can set $D_n(\cdot,\cdot) = 1$ whenever $\|\alpha \odot \tilde{F}_n\| = 0$ and let $\lambda(1) = 1$, and thus the foregoing arguments yield that Condition (11.8) implies manageability of the triangular array $\{f_{n1}(t), \ldots, f_{nn}(t), t \in T\}$.

Now the remaining conditions of the theorem can easily be show to imply Conditions (B) through (E) of Theorem 11.16 for the new triangular array and envelope vector $\tilde{F}_n$. Hence the desired results follow from Theorem 11.16.□

The following lemma gives us an important example of a sequence of classes $\mathcal{F}_n$ that satisfies Condition (11.8):

LEMMA 11.21 *For fixed index set $T$, let $\mathcal{F}_n = \{f_{n,t} : t \in T\}$ be a VC class of measurable functions with VC-index $V_n$ and integrable envelope $F_n$, for all $n \geq 1$, and assume $\sup_{n\geq 1} V_n = V < \infty$. Then the sequence $\mathcal{F}_n$ satisfies Condition (11.8).*

**Proof.** By Theorem 9.3, there exists a universal constant $K$ depending only on $V$ such that

$$N\left(\epsilon\|F_n\|_{Q,2}, \mathcal{F}_n, L_2(Q)\right) \leq K\left(\frac{1}{\epsilon}\right)^{2(V-1)},$$

for all $0 < \epsilon \leq 1$. Note that we have extended the range of $\epsilon$ to include 1, but this presents not difficulty since $N(\|F_n\|_{Q,2}, \mathcal{F}_n, L_2(Q)) = 1$ always holds by the definition of an envelope function. The desired result now follows since $V_n \leq V$ for all $n \geq 1$.□

As a simple example, let $T \subset \mathbb{R}$ and assume that $f_{n,t}(x)$ is always monotone increasing in $t$. Then $\mathcal{F}_n$ always has VC-index 2 by Lemma 9.10, and hence Lemma 11.21 applies. Thus (11.8) holds. This particular situation will apply later in Section 14.5.2 when we study the weak convergence of a certain monotone density estimator. This Condition (11.8) is quite similar to the BUEI condition for fixed function classes, and most of the preservation results of Section 9.1.2 will also apply. The following proposition, the proof of which is saved as an exercise, is one such preservation result:

PROPOSITION 11.22 *Let $\mathcal{G}_n$ and $\mathcal{H}_n$ be sequences of classes of measurable functions with respective envelope sequences $G_n$ and $H_n$, where Condition (11.8) is satisfied for the sequences $(\mathcal{F}_n, F_n) = (\mathcal{G}_n, G_n)$ and $(\mathcal{F}_n, F_n) = (\mathcal{H}_n, H_n)$. Then Condition (11.8) is also satisfied for the sequence of classes $\mathcal{F}_n = \mathcal{G}_n \cdot \mathcal{H}_n$ (consisting of all pairwise products) and the envelope sequence $F_n = G_n H_n$.*

We now present a weighted bootstrap result for this setting. Let $Z \equiv \{z_i, i \geq 1\}$ be a sequence of random variables satisfying

(F) The $\{z_i\}$ are positive, i.i.d. random variables which are independent of the data $X_1, X_2, \ldots$ and which have mean $0 < \mu < \infty$ and variance $0 < \tau^2 < \infty$.

The weighted bootstrapped process we propose for use here is

$$\hat{X}_n(t) \equiv \frac{\mu}{\tau} n^{-1/2} \sum_{i=1}^{n} \left(\frac{z_i}{\bar{z}_n} - 1\right) f_{n,t}(X_i),$$

where $\bar{z}_n \equiv n^{-1}\sum_{i=1}^{n} z_i$. The following theorem tells us that this is a valid bootstrap procedure:

THEOREM 11.23 *Suppose the class of functions $\mathcal{F}_n$, with envelope $F_n$, satisfies the conditions of Theorem 11.20 and the sequence $\{z_i, i \geq 1\}$ satisfies Condition (F) above. Then the conclusions of Theorem 11.20 obtain, $\hat{X}_n$ is asymptotically measurable, and $\hat{X}_n \overset{\mathrm{P}}{\underset{Z}{\rightsquigarrow}} X$.*

**Proof.** Note that

$$\begin{aligned}
\frac{\tau}{\mu}\hat{X}_n(t) &= n^{-1/2}\sum_{i=1}^{n}\left(\frac{z_i}{\bar{z}_n}-1\right)(f_{n,t}(X_i)-Pf_{n,t}) \\
&= n^{-1/2}\sum_{i=1}^{n}\left(\frac{z_i}{\mu}-1\right)(f_{n,t}(X_i)-Pf_{n,t}) \\
&\quad +n^{-1/2}\left(\frac{\mu}{\bar{z}_n}-1\right)\sum_{i=1}^{n}\left(\frac{z_i}{\mu}-1\right)(f_{n,t}(X_i)-Pf_{n,t}) \\
&\quad +n^{-1/2}\left(\frac{\mu}{\bar{z}_n}-1\right)\sum_{i=1}^{n}(f_{n,t}(X_i)-Pf_{n,t}) \\
&= A_n(t)+B_n(t)+C_n(t).
\end{aligned}$$

By Theorem 11.20, $\mathbb{G}_n f_{n,t} = O_P^{*T}(1)$, where $O_P^{*T}(1)$ denotes a term bounded in outer probability uniformly over $T$. Since Theorem 11.20 is really a special case of Theorem 11.16, we have by Theorem 11.19 that $n^{-1/2}\sum_{i=1}^{n}\frac{\mu}{\tau}\left(\frac{z_i}{\mu}-1\right)\times(f_{n,t}(X_i)-Pf_{n,t}) \overset{\mathrm{P}}{\underset{Z}{\rightsquigarrow}} X$, since $\frac{\mu}{\tau}\left(\frac{z_1}{\mu}-1\right)$ has mean zero and variance 1. Hence $(\mu/\tau)A_n$ is asymptotically measurable and $(\mu/\tau)A_n \overset{\mathrm{P}}{\underset{Z}{\rightsquigarrow}} X$. Moreover, since $\bar{z}_n \overset{\mathrm{as*}}{\to} \mu$, we now have that both $B_n = o_P^{*T}(1)$ and $C_n = o_P^{*T}(1)$, where $o_P^{*T}(1)$ denotes a term going to zero outer almost surely uniformly over $T$. The desired results now follow.□

## 11.6 Dependent Observations

In the section, we will review a number of empirical process results for dependent observations. A survey of recent results on this subject is *Empirical Process Techniques for Dependent Data*, edited by Dehling, Mikosch and Sørensen (2002); and a helpful general reference on theory for dependent observations is *Dependence in Probability and Statistics: A Survey of Recent Results*, edited by Eberlein and Taqqu (1986). Our focus here will be on strongly mixing stationary sequences (see Bradley, 1986). For the interested reader, a few results for non-stationary dependent sequences can be found in Andrews (1991), while several results for long range dependent sequences can be found in Dehling and Taqqu (1989), Yu (1994) and Wu (2003), among other references.

Let $X_1, X_2, \ldots$ be a *stationary* sequence of possibly dependent random variables on a probability space $(\Omega, \mathcal{D}, Q)$, and let $\mathcal{M}_a^b$ be the $\sigma$-field generated by $X_a, \ldots, X_b$. By stationary, we mean that for any set of positive integers $m_1, \ldots, m_k$, the joint distribution of $X_{m_1+j}, X_{m_2+j}, \ldots, X_{m_k+j}$ is unchanging for all integers $j \geq -m_1 + 1$. The sequence $\{X_i, i \geq 1\}$ is *strongly mixing* (also *$\alpha$-mixing*) if

$$\alpha(k) \equiv \sup_{m \geq 1} \left\{ |P(A \cap B) - P(A)P(B)| : A \in \mathcal{M}_1^m, B \in \mathcal{M}_{m+k}^\infty \right\} \to 0,$$

as $k \to \infty$, and it is *absolutely regular* (also *$\beta$-mixing*) if

$$\beta(k) \equiv \mathrm{E} \sup_{m \geq 1} \left\{ |P(B \,|\mathcal{M}_1^m) - P(B)| : B \in \mathcal{M}_{m+k}^\infty \right\} \to 0,$$

as $k \to \infty$. Other forms of mixing include $\rho$-mixing, $\phi$-mixing, $\psi$-mixing and $*$-mixing (see Definition 3.1 of Dehling and Philipp, 2002). Note that the stronger notion of *$m$-dependence*, where observations more than $m$ lags apart are independent, implies that $\beta(k) = 0$ for all $k > m$ and therefore also implies absolute regularity. It is also known that absolute regularity implies strong mixing (see Section 3.1 of Dehling and Philipp, 2002). Hereafter, we will restrict our attention to $\beta$-mixing sequences since these will be the most useful for our purposes.

We now present several empirical process Donsker and bootstrap results for absolutely regular stationary sequences. Let the values of $X_1$ lie in a Polish space $\mathcal{X}$ with distribution $P$, and let $\mathbb{G}_n$ be the empirical measure for the first $n$ observations of the sequence, i.e., $\mathbb{G}_n f = n^{1/2} \sum_{i=1}^n (f(X_i) - Pf)$, for any measurable $f : \mathcal{X} \mapsto \mathbb{R}$. We now present the following bracketing central limit theorem:

THEOREM 11.24 *Let $X_1, X_2, \ldots$ be a stationary sequence in a Polish space with marginal distribution $P$, and let $\mathcal{F}$ be a class of functions in $L_2(P)$. Suppose there exists a $2 < p < \infty$ such that*

*(a)* $\sum_{k=1}^\infty k^{2/(p-2)} \beta(k) < \infty$, *and*

*(b)* $J_{[]}(\infty, \mathcal{F}, L_p(P)) < \infty$.

*Then $\mathbb{G}_n \rightsquigarrow \mathbb{H}$ in $\ell^\infty(\mathcal{F})$, where $\mathbb{H}$ is a tight, mean zero Gaussian process with covariance*

$$\Gamma(f,g) \equiv \lim_{k \to \infty} \sum_{i=1}^\infty \mathrm{cov}(f(X_k), g(X_i)), \quad \textit{for all } f, g \in \mathcal{F}. \tag{11.9}$$

**Proof.** The result follows through Condition 2 of Theorem 5.2 of Dedecker and Louhichi (2002), after noting that their Condition 2 can be shown to be implied by Conditions (a) and (b) above, via arguments contained in Section 4.3 of Dedecker and Louhichi (2002). We omit the details.□

We next present a result for VC classes $\mathcal{F}$. In this case, we need to address the issue of measurability with some care. For what follows, let $\mathcal{B}$ be the $\sigma$-field of the measurable sets on $\mathcal{X}$. The class of functions $\mathcal{F}$ is *permissible* if it can be indexed by some set $T$, i.e., $\mathcal{F} = \{f(\cdot, t) : t \in T\}$ ($T$ could potentially be $\mathcal{F}$ for this purpose), in such a way that the following holds:

(a) $T$ is a Suslin metric space with Borel $\sigma$-field $\mathcal{B}(T)$,

(b) $f(\cdot,\cdot)$ is a $\mathcal{B}\times\mathcal{B}(T)$-measurable function from $\mathbb{R}\times T$ to $\mathbb{R}$.

Note that this definition is similar to the almost measurable Suslin condition of Section 11.5. We now have the following theorem:

THEOREM 11.25 *Let $X_1, X_2, \ldots$ be a stationary sequence in a Polish space with marginal distribution $P$, and let $\mathcal{F}$ be a class of functions in $L_2(P)$. Suppose there exists a $2 < p < \infty$ such that*

*(a)* $\lim_{k\to\infty} k^{2/(p-2)}(\log k)^{2(p-1)/(p-2)}\beta(k) = 0$, *and*

*(b) $\mathcal{F}$ is permissible, VC, and has envelope $F$ satisfying $P^*F^p < \infty$.*

*Then $\mathbb{G}_n \rightsquigarrow \mathbb{H}$ in $\ell^\infty(\mathcal{F})$, where $\mathbb{H}$ is as defined in Theorem 11.24 above.*

**Proof.** This theorem is essentially Theorem 2.1 of Arcones and Yu (1994), and the proof, under slightly different measurability conditions, can be found therein. The measurability issues, including the sufficiency of the permissibility condition, are addressed in the appendix of Yu (1994). We omit the details.□

Now we consider the bootstrap. A problem with the usual nonparametric bootstrap is that samples of $X_1, X_2, \ldots$ randomly drawn with replacement will be independent and will loose the dependency structure of the stationary sequence. Hence the usual bootstrap will generally not work. A modified bootstrap, the *moving blocks bootstrap* (MBB), was independently introduce by Künsch (1989) and Liu and Singh (1992) to address this problem. The method works as follows for a stationary sample $X_1, \ldots, X_n$: For a chosen block length $b \le n$, extend the sample by defining $X_{n+1}, \ldots, X_{n+b-1} = X_1, \ldots, X_b$ and let $k$ be the smallest integer such that $kb \ge n$. Now define blocks (as row vectors) $B_i = (X_i, X_{i+1}, \ldots, X_{i+b-1})$, for $i = 1, \ldots, n$, and sample from the $B_i$s with replacement to obtain $k$ blocks $B_1^*, B_2^*, \ldots, B_k^*$. The bootstrapped sample $X_1^*, \ldots, X_n^*$ consists of the first $n$ observations from the row vector $(B_1^*, \ldots, B_k^*)$. The bootstrapped empirical measure indexed by the class $\mathcal{F}$ is then defined as

$$\mathbb{G}_n^* f \equiv n^{-1/2} \sum_{i=1}^{n} (f(X_i^*) - \mathbb{P}_n f),$$

for all $f \in \mathcal{F}$, where $\mathbb{P}_n f \equiv n^{-1}\sum_{i=1}^n f(X_i)$ is the usual empirical probability measure (except that the data are now potentially dependent).

For now, we will assume that $X_1, X_2, \ldots$ are real-valued, although the results probably could be extended to general Polish-valued random variables. MBB bootstrap consistency has been established for bracketing entropy in Bühlmann (1995), although the entropy requirements are much stronger than those of Theorem 11.24 above, and also for VC classes in Radulović (1996). Other interesting, related references are Naik-Nimbalker and Rajarshi (1994) and Peligrad (1998), among others. We conclude this

section by presenting Radulović's (1996) Theorem 1 (slightly modified to address measurability) without proof:

THEOREM 11.26 *Let $X_1, X_2, \ldots$ be a stationary sequence of real random variables with marginal distribution $P$, and let $\mathcal{F}$ be a class of functions in $L_2(P)$. Also assume that $X_1^*, \ldots, X_n^*$ are generated by the MBB procedure with block size $b(n) \to \infty$, as $n \to \infty$, and that there exists $2 < p < \infty$, $q > p/(p-2)$, and $0 < \rho < (p-2)/[2(p-1)]$ such that*

*(a)* $\limsup_{k\to\infty} k^q \beta(k) < \infty$,

*(b)* $\mathcal{F}$ *is permissible, VC, and has envelope* $F$ *satisfying* $P^*F^p < \infty$, *and*

*(c)* $\limsup_{n\to\infty} n^{-\rho} b(n) < \infty$.

*Then* $\mathbb{G}_n^* \overset{\mathrm{P}}{\underset{*}{\rightsquigarrow}} \mathbb{H}$ *in* $\ell^\infty(\mathcal{F})$, *where* $\mathbb{H}$ *is as defined in Theorem 11.24 above.*

## 11.7 Proofs

**Proofs of Theorems 11.4 and 11.5.** Consider the class $\mathcal{F}' \equiv 1/2 + \mathcal{F}/(2M)$, and note that all functions $f \in \mathcal{F}'$ satisfy $0 \le f \le 1$. Moreover, $\|\mathbb{G}_n\|_{\mathcal{F}} = 2M\|\mathbb{G}_n\|_{\mathcal{F}'}$. Thus, if we prove the results for $\mathcal{F}'$, we are done.

For Theorem 11.4, the Condition (11.1) is also satisfied if we replace $\mathcal{F}$ with $\mathcal{F}'$. This can be done without changing $W$, although we may need to change $K$ to some $K' < \infty$. Theorem 2.14.10 of VW now yields that

$$\mathrm{P}^*(\|\mathbb{G}_n\|_{\mathcal{F}'} > t) \quad \le \quad Ce^{Dt^{U+\delta}} e^{-2t^2}, \tag{11.10}$$

for every $t > 0$ and $\delta > 0$, where $U = W(6-W)/(2+W)$ and $C$ and $D$ depend only on $K'$, $W$, and $\delta$. Since $0 < W < 2$, it can be shown that $0 < U < 2$. Accordingly, choose a $\delta > 0$ so that $U + \delta < 2$. Now, it can be shown that there exists a $C^* < \infty$ and $K^* > 0$ so that (11.10)$\le C^* e^{-K^* t^2}$, for every $t > 0$. Theorem 11.4 now follows by Lemma 8.1.

For Theorem 11.5, the Condition (11.2) implies the existence of a $K'$ so that $N_{[]}(\epsilon, \mathcal{F}', L_2(P)) \le (K'/\epsilon)^V$, for all $0 < \epsilon < K'$, where $V$ is the one in (11.2). Now Theorem 2.14.9 of VW yields that

$$\mathrm{P}^*(\|\mathbb{G}_n\|_{\mathcal{F}'} > t) \quad \le \quad Ct^V e^{-2t^2}, \tag{11.11}$$

for every $t > 0$, where the constant $C$ depends only on $K'$ and $V$. Thus there exists a $C^* < \infty$ and $K^* > 0$ such that (11.11)$\le C^* e^{-K^* t^2}$, for every $t > 0$. The desired result now follows.□

**Proof of Theorem 11.9.** For the proof of result (i), note that the cosines are bounded, and thus the series defining $\mathcal{F}_1$ is automatically pointwise convergent by the discussion prior to Theorem 11.9. Now, the Cramér-von Mises statistic is the square of the $L_2[0,1]$ norm of the function $t \mapsto$

$\mathbb{G}_n(t) \equiv \mathbb{G}_n 1\{X \le t\}$. Since the functions $\{g_i \equiv \sqrt{2}\sin \pi jt : j = 1, 2, \ldots\}$ form an orthonormal basis for $L_2[0,1]$, Parseval's formula tells us that the integral of the square of any function in $L_2[0,1]$ can be replaces by the sum of the squares of the Fourier coefficients. This yields:

$$\int_0^1 \mathbb{G}_n^2(t)dt = \sum_{i=1}^{\infty}\left[\int_0^1 \mathbb{G}_n(t)g_i(t)dt\right]^2 = \sum_{i=1}^{\infty}\left[\mathbb{G}_n\int_0^X g_i(t)dt\right]^2. \tag{11.12}$$

But since $\int_0^x g_i(t)dt = -(\pi i)^{-1}\sqrt{2}\cos \pi ix$, the last term in (11.12) becomes $\sum_{i=1}^{\infty}\mathbb{G}_n^2(\sqrt{2}\cos \pi iX)/(\pi i)^2$. Standard methods can now be used to establish that this converges weakly to the appropriate limiting distribution. An alternative proof can be obtained via the relation (11.3) and the fact that $\mathcal{F}_1$ is an elliptical class and hence Donsker. Now (i) follows.

For result (ii), note that the orthonormalized Legendre polynomials can be obtained by applying the Gram-Schmidt procedure to the functions 1, $u$, $u^2$, .... By problem 2.13.1 of VW, the orthonormalized Legendre polynomials satisfy the differential equations $(1-u^2)p_j''(u) - 2up_j'(u) = -j(j+1)p_j(u)$, for all $u \in [-1,1]$ and integers $j \ge 1$. By change of variables, followed by partial integration and use of this differential identity, we obtain

$$\begin{aligned}
2\int_0^1 & p_i'(2t-1)p_j'(2t-1)t(1-t)dt \\
&= \frac{1}{4}\int_{-1}^1 p_i'(u)p_j'(u)(1-u^2)du \\
&= -\frac{1}{4}\int_{-1}^1 p_i(u)\left[p_j''(u)(1-u^2) - 2up_j'(u)\right]du \\
&= \frac{1}{4}j(j+1)1\{i=j\}.
\end{aligned}$$

Thus the functions $2\sqrt{2}p_j'(2t-1)\sqrt{t(1-t)}/\sqrt{j(j+1)}$, with $j$ ranging over the positive integers, form an orthonormal basis for $L_2[0,1]$. By Parseval's formula, we obtain

$$\begin{aligned}
\int_0^1 \frac{\mathbb{G}_n^2(t)}{t(1-t)}dt &= \sum_{j=1}^{\infty}\left[\int_0^1 \mathbb{G}_n(t)p_i'(2t-1)dt\right]^2 \frac{8}{j(j+1)} \\
&= \sum_{j=1}^{\infty}\frac{2}{j(j+1)}\mathbb{G}_n^2\left(p_j(2t-1)\right).
\end{aligned}$$

By arguments similar to those used to establish (i), we can now verify that (ii) follows.□

**Proof of Corollary 11.10.** By transforming the data using the transform $x \mapsto F_0(x)$, we can, without loss of generality, assume that the data are all i.i.d. uniform and reduce the interval of integration to $[0,1]$. Now

Proposition 7.27 yields that $\hat{T}_1 \rightsquigarrow \int_0^1 \mathbb{G}^2(t)dt$, where $\mathbb{G}(t)$ is a standard Brownian bridge. Now Parseval's formula and arguments used in the proof of Theorem 11.9 yield that

$$\int_0^1 \mathbb{G}^2(t)dt = \sum_{i=1}^{\infty} \mathbb{G}^2(\sqrt{2}\cos\pi x)/(\pi i)^2 \equiv T_1^*,$$

where now $\mathbb{G}$ is a mean zero Gaussian Brownian bridge random measure, where the covariance between $\mathbb{G}(f)$ and $\mathbb{G}(g)$, where $f, g : [0,1] \mapsto \mathbb{R}$, is $\int_0^1 f(s)g(s)ds - \int_0^1 f(s)ds \int_0^1 g(s)ds$. The fact that $T_1^*$ is tight combined with the covariance structure of $\mathbb{G}$ yields that $T_1^*$ has the same distribution as $T_1$, and the desired weak convergence result for $\hat{T}_1$ follows.

For $\hat{T}_2$, apply the same transformation as above so that, without loss of generality, we can again assume that the data are i.i.d. uniform and that the interval of integration is $[0,1]$. Let

$$\tilde{\mathbb{G}}_n \equiv \sqrt{\frac{n_1 n_2}{n_1+n_2}}\left(\hat{F}_{n,1} - \hat{F}_{n,2}\right),$$

and fix $\epsilon \in (0, 1/2)$. We can now apply Proposition 7.27 to verify that

$$\int_\epsilon^{1-\epsilon} \frac{\tilde{\mathbb{G}}_n^2(s)}{\hat{F}_{n,0}(s)\left[1-\hat{F}_{n,0}(s)\right]} d\hat{F}_{n,0}(s) \rightsquigarrow \int_\epsilon^{1-\epsilon} \frac{\mathbb{G}^2(s)}{s(1-s)}ds.$$

Note also that Fubini's theorem yields that both $\mathrm{E}\left\{\int_0^\epsilon \mathbb{G}^2(s)/[s(1-s)]ds\right\}$ $= \epsilon$ and $\mathrm{E}\left\{\int_{1-\epsilon}^1 \mathbb{G}^2(s)/[s(1-s)]ds\right\} = \epsilon$.

We will now work towards bounding $\int_0^\epsilon \left(\tilde{\mathbb{G}}_n(s)/\left\{\hat{F}_{n,0}(s)[1-\hat{F}_{n,0}(s)]\right\}\right) ds$. Fix $s \in (0,\epsilon)$ and note that, under the null hypothesis, the conditional distribution of $\tilde{\mathbb{G}}_n(s)$ given $\hat{F}_{n,0}(s) = m$ has the form

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}\left(\frac{A}{n_1} - \frac{m-A}{n_2}\right),$$

where $A$ is hypergeometric with density

$$\mathrm{P}(A=a) = \left(\begin{array}{c} n_1 \\ a \end{array}\right)\left(\begin{array}{c} n_1 \\ m-a \end{array}\right) / \left(\begin{array}{c} n \\ m \end{array}\right),$$

where $a$ is any integer between $(m-n_2)\vee 0$ and $m \wedge n_1$. Hence

$$\begin{aligned} \mathrm{E}\left[\tilde{\mathbb{G}}_n^2(s)\middle|\, \hat{F}_{n,0}(s) = m\right] &= \frac{n_1+n_2}{n_1 n_2}\mathrm{var}(A) \\ &= \frac{n}{n_1 n_2}\left(m\frac{n_1 n_2(n-m)}{n^2(n-1)}\right) \\ &= \frac{n}{n-1}\left[\frac{m(n-m)}{n^2}\right]. \end{aligned}$$

Thus

$$\mathrm{E}\left[\int_0^{\epsilon} \frac{\tilde{\mathbb{G}}_n^2(s)}{\hat{F}_{n,0}(s)[1-\hat{F}_{n,0}(s)]} ds\right] = \frac{n}{n-1}\mathrm{E}\hat{F}_{n,0}(\epsilon) \leq 2\epsilon,$$

for all $n \geq 2$. Similar arguments verify that

$$\mathrm{E}\left[\int_{1-\epsilon}^{1} \frac{\tilde{\mathbb{G}}_n^2(s)}{\hat{F}_{n,0}(s)[1-\hat{F}_{n,0}(s)]} ds\right] \leq 2\epsilon,$$

for all $n \geq 2$. Since $\epsilon$ was arbitrary, we now have that

$$\hat{T}_2 \rightsquigarrow \int_0^1 \frac{\mathbb{G}^2(s)}{s(1-s)} ds \equiv T_2^*,$$

where $\mathbb{G}$ is the same Brownian bridge process used in defining $T_1^*$. Now we can again use arguments from the proof of Theorem 11.9 to obtain that $T_2^*$ has the same distribution as $T_2$.□

## 11.8 Exercises

**11.8.1.** Verify that $F^*$ in the proof of Theorem 11.6 is continuous.

**11.8.2.** Show that when $P_n$ and $P$ satisfy (11.4), we have that $P_n h \to Ph$, as $n \to \infty$, for all bounded and measurable $h$.

**11.8.3.** Prove Proposition 11.22. Hint: Consider the arguments used in the proof of Theorem 9.15.

## 11.9 Notes

Theorems 11.4 and 11.5 are inspired by Theorem 2.14.9 of VW, and Theorem 11.7 is derived from Theorem 2.13.1 of VW. Theorem 11.8 is Theorem 2.13.2 of VW, while Theorem 11.9 is derived from Examples 2.13.3 and 2.13.5 of VW. Much of the structure of Section 11.4 comes from Kosorok (2003), although Theorem 11.17 was derived from material in Section 5 of Pollard (1990). Lemma 11.15 and Theorems 11.16, 11.18 and 11.19, are Lemma 2 and Theorems 1 (with a minor modification), 3 and 2, respectively, of Kosorok (2003).

# 12
# The Functional Delta Method

In this chapter, we build on the presentation of the functional delta method given in Section 2.2.4. Recall the concept of Hadamard differentiability introduced in this section and also defined more precisely in Section 6.3. The key result of Section 2.2.4 is that the delta method and its bootstrap counterpart work provided the map $\phi$ is Hadamard differentiable tangentially to a suitable set $\mathbb{D}_0$. We first present in Section 12.1 clarifications and proofs of the two main theorems given in Section 2.2.4, the functional delta method for Hadamard differentiable maps (Theorem 2.8 on Page 22) and the conditional analog for the bootstrap (Theorem 2.9 on Page 23). We then give in Section 12.2 several important examples of Hadamard differentiable maps of use in statistics, along with specific illustrations of how those maps are utilized.

## 12.1 Main Results and Proofs

In this section, we first prove the functional delta method theorem (Theorem 2.8 on Page 22) and then restate and prove Theorem 2.9 from Page 23. Before proceeding, recall that $X_n$ in the statement of Theorem 2.8 is a random quantity that takes its values in a normed space $\mathbb{D}$.

**Proof of Theorem 2.8 (Page 22).** Consider the map $h \mapsto r_n(\phi(\theta + r_n^{-1}h) - \phi(\theta)) \equiv g_n(h)$, and note that it is defined on the domain $\mathbb{D}_n \equiv \{h : \theta + r_n^{-1}h \in \mathbb{D}_\phi\}$ and satisfies $g_n(h_n) \to \phi'_\theta(h)$ for every $h_n \to h \in \mathbb{D}_0$ with $h_n \in \mathbb{D}_n$. Thus the conditions of the extended continuous mapping

theorem (Theorem 7.24) are satisfied by $g(\cdot) = \phi'_\theta(\cdot)$. Hence conclusion (i) of that theorem implies $g_n(r_n(X_n - \theta)) \rightsquigarrow \phi'_\theta(X)$.□

We now restate and prove Theorem 2.9 on Page 23. The restatement clarifies the measurability condition. Before proceeding, recall the definitions of $\mathbb{X}_n$ and $\hat{\mathbb{X}}_n$ in the statement of Theorem 2.9. Specifically, $\mathbb{X}_n(X_n)$ is a sequence of random elements in a normed space $\mathbb{D}$ based on the data sequence $\{X_n, n \geq 1\}$, while $\hat{\mathbb{X}}_n(X_n, W_n)$ is a bootstrapped version of $\mathbb{X}_n$ based on both the data sequence and a sequence of weights $W = \{W_n, n \geq 1\}$. Note that the proof of this theorem utilizes the bootstrap continuous mapping theorem (Theorem 10.8). Here is the restated version of Theorem 2.9 from Page 23:

THEOREM 12.1 *For normed spaces $\mathbb{D}$ and $\mathbb{E}$, let $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ be Hadamard-differentiable at $\mu$ tangentially to $\mathbb{D}_0 \subset \mathbb{D}$, with derivative $\phi'_\mu$. Let $\mathbb{X}_n$ and $\hat{\mathbb{X}}_n$ have values in $\mathbb{D}_\phi$, with $r_n(\mathbb{X}_n - \mu) \rightsquigarrow \mathbb{X}$, where $\mathbb{X}$ is tight and takes its values in $\mathbb{D}_0$ for some sequence of constants $0 < r_n \to \infty$, the maps $W_n \mapsto h(\hat{\mathbb{X}}_n)$ are measurable for every $h \in C_b(\mathbb{D})$ outer almost surely, and where $r_n c(\hat{\mathbb{X}}_n - \mathbb{X}_n) \underset{W}{\overset{\mathrm{P}}{\rightsquigarrow}} \mathbb{X}$, for a constant $0 < c < \infty$. Then $r_n c(\phi(\hat{\mathbb{X}}_n) - \phi(\mathbb{X}_n)) \underset{W}{\overset{\mathrm{P}}{\rightsquigarrow}} \phi'_\mu(\mathbb{X})$.*

**Proof.** We can, without loss of generality, assume that $\mathbb{D}_0$ is complete and separable (since $\mathbb{X}$ is tight), that $\mathbb{D}$ and $\mathbb{E}$ are both complete, and that $\phi'_\mu : \mathbb{D} \mapsto \mathbb{E}$ is continuous on all of $\mathbb{D}$, although it is permitted to not be bounded or linear off of $\mathbb{D}_0$. To accomplish this, one can apply the Dugundji extension theorem (Theorem 10.9) which extends any continuous operator defined on a closed subset to the entire space. It may be necessary to replace $\mathbb{E}$ with its closed linear span to accomplish this.

We can now use arguments nearly identical to those used in the proof given in Section 10.1 of Theorem 10.4 to verify that, unconditionally, both $\hat{\mathbb{U}}_n \equiv r_n(\hat{\mathbb{X}}_n - \mathbb{X}_n) \rightsquigarrow c^{-1}\mathbb{X}$ and $r_n(\hat{\mathbb{X}}_n - \mu) \rightsquigarrow Z$, where $Z$ is a tight random element. Fix some $h \in BL_1(\mathbb{D})$, define $\mathbb{U}_n \equiv r_n(\mathbb{X}_n - \mu)$, and let $\tilde{\mathbb{X}}_1$ and $\tilde{\mathbb{X}}_2$ be two independent copies of $\mathbb{X}$. We now have that

$$\begin{aligned}
&\left|\mathrm{E}^* h(\hat{\mathbb{U}}_n + \mathbb{U}_n) - \mathrm{E} h(c^{-1}\tilde{\mathbb{X}}_1 + \tilde{\mathbb{X}}_2)\right| \\
&\qquad \leq \left|\mathrm{E}_{X_n}\mathrm{E}_{W_n} h(\hat{\mathbb{U}}_n + \mathbb{U}_n)^* - \mathrm{E}^*\mathrm{E}_{W_n} h(\hat{\mathbb{U}}_n + \mathbb{U}_n)\right| \\
&\qquad\quad + \mathrm{E}^*\left|\mathrm{E}_{W_n} h(\hat{\mathbb{U}}_n + \mathbb{U}_n) - \mathrm{E}_{\tilde{\mathbb{X}}_1} h(c^{-1}\tilde{\mathbb{X}}_1 + \mathbb{U}_n)\right| \\
&\qquad\quad + \left|\mathrm{E}^*\mathrm{E}_{\tilde{\mathbb{X}}_1} h(c^{-1}\tilde{\mathbb{X}}_1 + \mathbb{U}_n) - \mathrm{E}_{\tilde{\mathbb{X}}_2}\mathrm{E}_{\tilde{\mathbb{X}}_1} h(c^{-1}\tilde{\mathbb{X}}_1 + \tilde{\mathbb{X}}_2)\right|,
\end{aligned}$$

where $\mathrm{E}_{W_n}$, $\mathrm{E}_{\tilde{\mathbb{X}}_1}$ and $\mathrm{E}_{\tilde{\mathbb{X}}_2}$ are expectations taken over the bootstrap weights, $\tilde{\mathbb{X}}_1$ and $\tilde{\mathbb{X}}_2$, respectively. The first term on the right in the above expression goes to zero by the asymptotic measurability of $\hat{\mathbb{U}}_n + \mathbb{U}_n = r_n(\hat{\mathbb{X}}_n - \mu)$. The second term goes to zero by the fact that $\hat{\mathbb{U}}_n \underset{W}{\overset{\mathrm{P}}{\rightsquigarrow}} c^{-1}\mathbb{X}$ combined with

the fact that the map $x \mapsto h(x+\mathbb{U}_n)$ is Lipschitz continuous with Lipschitz constant 1 outer almost surely. The third term goes to zero since $\mathbb{U}_n \rightsquigarrow \mathbb{X}$ and the map $x \mapsto \mathrm{E}_{\tilde{\mathbb{X}}_1} h(\tilde{\mathbb{X}}_1 + x)$ is also Lipschitz continuous with Lipschitz constant 1. Since $h$ was arbitrary, we have by the Portmanteau theorem that, unconditionally,

$$r_n \begin{pmatrix} \hat{\mathbb{X}}_n - \mu \\ \mathbb{X}_n - \mu \end{pmatrix} \rightsquigarrow \begin{pmatrix} c^{-1}\tilde{\mathbb{X}}_1 + \tilde{\mathbb{X}}_2 \\ \tilde{\mathbb{X}}_2 \end{pmatrix}.$$

Now the functional delta method (Theorem 2.8) yields

$$r_n \begin{pmatrix} \phi(\hat{\mathbb{X}}_n) - \phi(\mu) \\ \phi(\mathbb{X}_n) - \phi(\mu) \\ \hat{\mathbb{X}}_n - \mu \\ \mathbb{X}_n - \mu \end{pmatrix} \rightsquigarrow \begin{pmatrix} \phi'_\mu(c^{-1}\tilde{\mathbb{X}}_1 + \mathbb{X}_2) \\ \phi'_\mu(\mathbb{X}_2) \\ c^{-1}\tilde{\mathbb{X}}_1 + \mathbb{X}_2 \\ \mathbb{X}_2 \end{pmatrix},$$

since the map $(x, y) \mapsto (\phi(x), \phi(y), x, y)$ is Hadamard differentiable at $(\mu, \mu)$ tangentially to $\mathbb{D}_0$. This implies two things. First,

$$r_n c \begin{pmatrix} \phi(\hat{\mathbb{X}}_n) - \phi(\mathbb{X}_n) \\ \hat{\mathbb{X}}_n - \mathbb{X}_n \end{pmatrix} \rightsquigarrow \begin{pmatrix} \phi'_\mu(\mathbb{X}) \\ \mathbb{X} \end{pmatrix},$$

since $\phi'_\mu$ is linear on $\mathbb{D}_0$. Second, the usual continuous mapping theorem now yields that, unconditionally,

$$r_n c(\phi(\hat{\mathbb{X}}_n) - \phi(\mathbb{X}_n)) - \phi'_\mu(r_n c(\hat{\mathbb{X}}_n - \mathbb{X}_n)) \xrightarrow{\mathrm{P}} 0, \tag{12.1}$$

since the map $(x, y) \mapsto x - \phi'_\mu(y)$ is continuous on all of $\mathbb{E} \times \mathbb{D}$.

Now for any map $h \in C_b(\mathbb{D})$, the map $x \mapsto h(r_n c(x - \mathbb{X}_n))$ is continuous and bounded for all $x \in \mathbb{D}$ outer almost surely. Thus the maps $W_n \mapsto h(r_n c(\hat{\mathbb{X}}_n - \mathbb{X}_n))$ are measurable for every $h \in C_b(\mathbb{D})$ outer almost surely. Hence the bootstrap continuous mapping theorem, Theorem 10.8, yields that $\phi'_\mu(r_n c(\hat{\mathbb{X}}_n - \mathbb{X}_n)) \overset{\mathrm{P}}{\underset{W}{\rightsquigarrow}} \phi'_\mu(\mathbb{X})$. The desired result now follows from (12.1).□

## 12.2 Examples

We now give several important examples of Hadamard differentiable maps, along with illustrations of how these maps are utilized in statistical applications.

### *12.2.1 Composition*

Recall from Section 2.2.4 the map $\phi : \mathbb{D}_\phi \mapsto D[0,1]$, where $\phi(f) = 1/f$ and $\mathbb{D}_\phi = \{f \in D[0,1] : |f| > 0\}$. In that section, we established that $\phi$ was

Hadamard differentiable, tangentially to $D[0,1]$, with derivative at $\theta \in \mathbb{D}_\phi$ equal to $h \mapsto -h/\theta^2$. This is a simple example of the following general composition result:

LEMMA 12.2 *Let $g : B \subset \bar{\mathbb{R}} \equiv [-\infty, \infty] \mapsto \mathbb{R}$ be differentiable with derivative continuous on all closed subsets of $B$, and let $\mathbb{D}_\phi = \{A \in \ell^\infty(\mathcal{X}) : \{R(A)\}^\delta \subset B$ for some $\delta > 0\}$, where $\mathcal{X}$ is a set, $R(C)$ denotes the range of the function $C \in \ell^\infty(\mathcal{X})$, and $D^\delta$ is the $\delta$-enlargement of the set $D$. Then $A \mapsto g \circ A$ is Hadamard-differentiable as a map from $\mathbb{D}_\phi \subset \ell^\infty(\mathcal{X})$ to $\ell^\infty(\mathcal{X})$, at every $A \in \mathbb{D}_\phi$. The derivative is given by $\phi'_A(\alpha) = g'(A)\alpha$, where $g'$ is the derivative of $g$.*

Before giving the proof, we briefly return to our simple example of the reciprocal map $A \mapsto 1/A$. The differentiability of this map easily generalizes from $D[0,1]$ to $\ell^\infty(\mathcal{X})$, for arbitrary $\mathcal{X}$, provided we restrict the domain of the reciprocal map to $\mathbb{D}_\phi = \{A \in \ell^\infty(\mathcal{X}) : \inf_{x\in\mathcal{X}} |A(x)| > 0\}$. This follows after applying Lemma 12.2 to the set $B = [-\infty, 0) \cup (0, \infty]$.

**Proof of Lemma 12.2.** Note that $\mathbb{D} = \ell^\infty(\mathcal{X})$ in this case, and that the tangent set for the derivative is all of $\mathbb{D}$. Let $t_n$ be any real sequence with $t_n \to 0$, let $\{h_n\} \in \ell^\infty(\mathcal{X})$ be any sequence converging to $\alpha \in \ell^\infty(\mathcal{X})$ uniformly, and define $A_n = A + t_n h_n$. Then, by the conditions of the theorem, there exists a closed $B_1 \subset B$ such that $\{R(A) \cup R(A_n)\}^\delta \subset B_1$ for some $\delta > 0$ and all $n$ large enough. Hence

$$\sup_{x\in\mathcal{X}} \left| \frac{g(A(x) + t_n h_n(x)) - g(A(x))}{t_n} - g'(A(x))\alpha(x) \right| \to 0,$$

as $n \to \infty$, since continuous functions on closed sets are bounded and thus $g'$ is uniformly continuous on $B_1$.□

### *12.2.2 Integration*

For an $M < \infty$ and an interval $[a,b] \in \bar{\mathbb{R}}$, let $BV_M[a,b]$ be the set of all functions $A \in D[a,b]$ with total variation $|A(0)| + \int_{(a,b]} |dA(s)| \le M$. In this section, we consider, for given functions $A \in D[a,b]$ and $B \in BV_M[a,b]$ and domain $\mathbb{D}_M \equiv D[a,b] \times BV_M[a,b]$, the maps $\phi : \mathbb{D}_M \mapsto \mathbb{R}$ and $\psi : \mathbb{D}_M \mapsto D[a,b]$ defined by

$$(12.2)\quad \phi(A,B) = \int_{(a,b]} A(s)dB(s) \quad \text{and} \quad \psi(A,B)(t) = \int_{(a,t]} A(s)dB(s).$$

The following lemma verifies that these two maps are Hadamard differentiable:

LEMMA 12.3 *For each fixed $M < \infty$, the maps $\phi : \mathbb{D}_M \mapsto \mathbb{R}$ and $\psi : \mathbb{D}_M \mapsto D[a,b]$ defined in (12.2) are Hadamard differentiable at each $(A,B) \in \mathbb{D}_M$ with $\int_{(a,b]} |dA| < \infty$. The derivatives are given by*

$$\begin{aligned}\phi'_{A,B}(\alpha,\beta) &= \int_{(a,b]} Ad\beta + \int_{(a,b]} \alpha dB, \quad \textit{and} \\ \psi'_{A,B}(\alpha,\beta)(t) &= \int_{(a,t]} Ad\beta + \int_{(a,t]} \alpha dB.\end{aligned}$$

Note that in the above lemma we define $\int_{(a,t]} Ad\beta = A(t)\beta(t) - A(a)\beta(a) - \int_{(a,t]} \beta(s-)dA(s)$ so that the integral is well defined even when $\beta$ does not have bounded variation. We will present the proof of this lemma at the end of this section.

We now look at two statistical applications of Lemma 12.3, the two-sample Wilcoxon rank sum statistic, and the Nelson-Aalen integrated hazard estimator. Consider first the Wilcoxon statistic. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent samples from distributions $F$ and $G$ on the reals. If $\mathbb{F}_m$ and $\mathbb{G}_n$ are the respective empirical distribution functions, the Wilcoxon rank sum statistic for comparing $F$ and $G$ has the form

$$T_1 = m \int_{\mathbb{R}} (m\mathbb{F}_m(x) + n\mathbb{G}_n(x)) d\mathbb{F}_m(x).$$

If we temporarily assume that $F$ and $G$ are continuous, then

$$\begin{aligned}T_1 &= mn \int_{\mathbb{R}} \mathbb{G}_n(x) d\mathbb{F}_m(x) + m^2 \int_{\mathbb{R}} \mathbb{F}_m(x) d\mathbb{F}_m(x) \\ &= mn \int_{\mathbb{R}} \mathbb{G}_n(x) d\mathbb{F}_m(x) + \frac{m^2+m}{2} \\ &\equiv mnT_2 + \frac{m^2+m}{2},\end{aligned}$$

where $T_2$ is the Mann-Whitney statistic. When $F$ or $G$ have atoms, the relationship between the Wilcoxon and Mann-Whitney statistics is more complex. We will now study the asymptotic properties of the Mann-Whitney version of the rank sum statistic, $T_2$.

For arbitrary $F$ and $G$, $T_2 = \phi(\mathbb{G}_n, \mathbb{F}_m)$, where $\phi$ is as defined in Lemma 12.3. Note that $F$, $G$, $\mathbb{F}_m$ and $\mathbb{G}_n$ all have total variation $\leq 1$. Thus Lemma 12.3 applies, and we obtain that the Hadamard derivative of $\phi$ at $(A,B) = (G,F)$ is the map $\phi'_{G,F}(\alpha,\beta) = \int_{\mathbb{R}} Gd\beta + \int_{\mathbb{R}} \alpha dF$, which is continuous and linear over $\alpha, \beta \in D[-\infty,\infty]$. If we assume that $m/(m+n) \to \lambda \in [0,1]$, as $m \wedge n \to \infty$, then

$$\sqrt{\frac{mn}{m+n}} \begin{pmatrix} \mathbb{G}_n - G \\ \mathbb{F}_m - F \end{pmatrix} \rightsquigarrow \begin{pmatrix} \sqrt{\lambda}\,\mathbb{B}_1(G) \\ \sqrt{1-\lambda}\,\mathbb{B}_2(F) \end{pmatrix},$$

where $\mathbb{B}_1$ and $\mathbb{B}_2$ are independent standard Brownian bridges. Hence $\mathbb{G}_G(\cdot) \equiv \mathbb{B}_1(G(\cdot))$ and $\mathbb{G}_F(\cdot) \equiv \mathbb{B}_2(F(\cdot))$ both live in $D[-\infty,\infty]$. Now Theorem 2.8 yields

$$T_2 \rightsquigarrow \sqrt{\lambda}\int_{\mathbb{R}} G d\mathbb{G}_F + \sqrt{1-\lambda}\int_{\mathbb{R}} \mathbb{G}_F dG,$$

as $m \wedge n \to \infty$. When $F = G$ and $F$ is continuous, this limiting distribution is mean zero normal with variance $1/12$. The delta method bootstrap, Theorem 12.1, is also applicable and can be used to obtain an estimate of the limiting distribution under more general hypotheses on $F$ and $G$.

We now shift our attention to the Nelson-Aalen estimator under right censoring. In the right censored survival data setting, an observation consists of the pair $(X, \delta)$, where $X = T \wedge C$ is the minimum of a failure time $T$ and censoring time $C$, and $\delta = 1\{T \leq C\}$. $T$ and $C$ are assumed to be independent. Let $F$ be the distribution function for $T$, and define the integrated baseline hazard for $F$ to be $\Lambda(t) = \int_0^t dF(s)/S(s-)$, where $S \equiv 1-F$ is the survival function. The Nelson-Aalen estimator for $\Lambda$, based on the i.i.d. sample $(X_1, \delta_1), \ldots, (X_n, \delta_n)$, is

$$\hat{\Lambda}_n(t) \equiv \int_{[0,t]} \frac{d\hat{N}_n(s)}{\hat{Y}_n(s)},$$

where $\hat{N}_n(t) \equiv n^{-1}\sum_{i=1}^n \delta_i 1\{X_i \leq t\}$ and $\hat{Y}_n(t) \equiv n^{-1}\sum_{i=1}^n 1\{X_i \geq t\}$. It is easy to verify that the classes $\{\delta 1\{X \leq t\}, t \geq 0\}$ and $\{1\{X \geq t\} : t \geq 0\}$ are both Donsker and hence that

$$\sqrt{n}\begin{pmatrix} \hat{N}_n - N_0 \\ \hat{Y}_n - Y_0 \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{G}_1 \\ \mathbb{G}_2 \end{pmatrix}, \tag{12.3}$$

where $N_0(t) \equiv P(T \leq t, C \geq T)$, $Y_0(t) \equiv P(X \geq t)$, and $\mathbb{G}_1$ and $\mathbb{G}_2$ are tight Gaussian processes with respective covariances $N_0(s \wedge t) - N_0(s)N_0(t)$ and $Y_0(s \vee t) - Y_0(s)Y_0(t)$ and with cross-covariance $1\{s \geq t\}[N_0(s) - N_0(t-)] - N_0(s)Y_0(t)$. Note that while we have already seen this survival set-up several times (eg., Sections 2.2.5 and 4.2.2), we are choosing to use slightly different notation than previously used to emphasize certain features of the underlying empirical processes.

The Nelson-Aalen estimator depends on the pair $(\hat{N}_n, \hat{Y}_n)$ through the two maps

$$(A, B) \mapsto \left(A, \frac{1}{B}\right) \mapsto \int_{[0,t]} \frac{1}{B} dA.$$

From Section 12.1.1, Lemma 12.3, and the chain rule (Lemma 6.19), it is easy to see that this composition map is Hadamard differentiable on a domain of the type $\{(A, B) : \int_{[0,\tau]} |dA(t)| \leq M, \inf_{t\in[0,\tau]} |B(t)| \geq \epsilon\}$ for a given $M < \infty$ and $\epsilon > 0$, at every point $(A, B)$ such that $1/B$ has bounded variation. Note that the interval of integration we are using, $[0, \tau]$, is left-closed rather than left-open as in the definition of $\psi$ given in (12.2). However, if we pick some $\eta > 0$, then in fact integrals over $[0, t]$, for any $t > 0$, of functions which have zero variation over $(-\infty, 0)$ are unchanged if

we replace the interval of integration with $(-\eta, t]$. Thus we will still be able to utilize Lemma 12.3 in our current set-up. In this case, the point $(A, B)$ of interest is $A = N_0$ and $B = Y_0$. Thus if $t$ is restricted to the interval $[0, \tau]$, where $\tau$ satisfied $Y_0(\tau) > 0$, then it is easy to see that the pair $(\hat{N}_n, \hat{Y}_n)$ is contained in the given domain with probability tending to 1 as $n \to \infty$. The derivative of the composition map is given by

$$(\alpha, \beta) \mapsto \left(\alpha, \frac{-\beta}{Y_0^2}\right) \mapsto \int_{[0,t]} \frac{d\alpha}{Y_0} - \int_{[0,t]} \frac{\beta dN_0}{Y_0^2}.$$

Thus from (12.3), we obtain via Theorem 2.8 that

$$\sqrt{n}(\hat{\Lambda}_n - \Lambda) \quad \rightsquigarrow \quad \int_{[0,(\cdot)]} \frac{d\mathbb{G}_1}{Y_0} - \int_{[0,(\cdot)]} \frac{\mathbb{G}_2 dN_0}{Y_0^2}. \tag{12.4}$$

The Gaussian process on the right side of (12.4) is equal to $\int_{[0,(\cdot)]} d\mathbb{M}/Y_0$, where $\mathbb{M}(t) \equiv \mathbb{G}_1(t) - \int_{[0,t]} \mathbb{G}_2 d\Lambda$ can be shown to be a Gaussian martingale with independent increments and covariance $\int_{[0,s\wedge t]}(1-\Delta\Lambda)d\Lambda$, where $\Delta A(t) \equiv A(t) - A(t-)$ is the mass at $t$ of a signed-measure $A$. This means that the Gaussian process on the right side of (12.4) is also a Gaussian martingale with independent increments but with covariance $\int_{[0,s\wedge t]}(1 - \Delta\Lambda)d\Lambda/Y_0$. A useful discussion of continuous time martingales arising in right censored survival data can be found in Fleming and Harrington (1991).

The delta method bootstrap, Theorem 12.1, is also applicable here and can be used to obtain an estimate of the limiting distribution. However, when $\Lambda$ is continuous over $[0, \tau]$, the independent increments structure implies that the limiting distribution is time-transformed Brownian motion. More precisely, the limiting process can be expressed as $\mathbb{W}(v(t))$, where $\mathbb{W}$ is standard Brownian motion on $[0, \infty)$ and $v(t) \equiv \int_{(0,t]} d\Lambda/Y_0$. As discussed in Chapter 7 of Fleming and Harrington (1991), this fact can be used to compute asymptotically exact simultaneous confidence bands for $\Lambda$.

**Proof of Lemma 12.3.** For sequences $t_n \to 0$, $\alpha_n \to \alpha$, and $\beta_n \to \beta$, define $A_n \equiv A + t_n\alpha_n$ and $B_n \equiv B + t_n\beta_n$. Since we require that $(A_n, B_n) \in \mathbb{D}_M$, we know that the total variation of $B_n$ is bounded by $M$. Consider first the derivative of $\psi$, and note that

$$\frac{\int_{(a,t]} A_n dB_n - \int_{(a,t]} A dB}{t_n} - \psi'_{A,B}(\alpha_n, \beta_n) \quad = \quad \int_{(a,t]} \alpha d(B_n - B) + \int_{(a,t]} (\alpha_n - \alpha) d(B_n - B). \tag{12.5}$$

Since it is easy to verify that the map $(\alpha, \beta) \mapsto \psi'_{A,B}(\alpha, \beta)$ is continuous and linear, the desired Hadamard differentiability of $\psi$ will follow provided

the right side of (12.5) goes to zero. To begin with, the second term on the right side goes to zero uniformly over $t \in (a, b]$, since both $B_n$ and $B$ have total variation bounded by $M$.

Now, for the first term on the right side of (12.5), fix $\epsilon > 0$. Since $\alpha$ is cadlag, there exists a partition $a = t_0 < t_1 < \cdots < t_m = b$ such that $\alpha$ varies less than $\epsilon$ over each interval $[t_{i-1}, t_i)$, $1 \leq i \leq m$, and $m < \infty$. Now define the function $\tilde{\alpha}$ to be equal to $\alpha(t_{i-1})$ over the interval $[t_{i-1}, t_i)$, $1 \leq i \leq m$, with $\tilde{\alpha}(b) = \alpha(b)$. Thus

$$\begin{aligned}
&\left\| \int_{(a,t]} \alpha d(B_n - B) \right\|_\infty \\
&\quad \leq \left\| \int_{(a,t]} (\alpha - \tilde{\alpha}) d(B_n - B) \right\|_\infty + \left\| \int_{(a,t]} \tilde{\alpha} d(B_n - B) \right\|_\infty \\
&\quad \leq \|\alpha - \tilde{\alpha}\|_\infty 2M + \sum_{i=1}^{m} |\alpha(t_{i-1})| \times |(B_n - B)(t_i) - (B_n - B)(t_{i-1})| \\
&\qquad + |\alpha(b)| \times |(B_n - B)(b)| \\
&\quad \leq \epsilon 2M + (2m+1)\|B_n - B\|_\infty \|\alpha\|_\infty \\
&\quad \to \epsilon 2M,
\end{aligned}$$

as $n \to \infty$. Since $\epsilon$ was arbitrary, we have that the first term on the right side of (12.5) goes to zero, as $n \to \infty$, and the desired Hadamard differentiability of $\psi$ follows.

Now the desired Hadamard differentiability of $\phi$ follows from the trivial but useful Lemma 12.4 below, by taking the extraction map $f : D[a, b] \mapsto \mathbb{R}$ defined by $f(x) = x(b)$, noting that $\phi = f(\psi)$, and then applying the chain rule for Hadamard derivatives (Lemma 6.19).□

LEMMA 12.4 *Let $T$ be a set and fix $T_0 \subset T$. Define the extraction map $f : \ell^\infty(T) \mapsto \ell^\infty(T_0)$ as $f(x) = \{x(t) : t \in T_0\}$. Then $f$ is Hadamard differentiable at all $x \in \ell^\infty(T)$ with derivative $f'_x(h) = \{h(t) : t \in T_0\}$.*

**Proof.** Let $t_n$ be any real sequence with $t_n \to 0$, and let $\{h_n\} \in \ell^\infty(T)$ be any sequence converging to $h \in \ell^\infty(T)$. The desired conclusion follows after noting that $t_n^{-1}[f(x + t_n h_n) - f(x)] = \{h_n(t) : t \in T_0\} \to \{h(t) : t \in T_0\}$, as $n \to \infty$.□

### *12.2.3 Product Integration*

For a function $A \in D(0, b]$, let $A^c(t) \equiv A(t) - \sum_{0<s\leq t} \Delta A(s)$, where $\Delta A$ is as defined in the previous section, be the continuous part of $A$. We define the product integral to be the map $A \mapsto \phi(A)$, where

$$\phi(A)(t) \equiv \prod_{0<s\leq t} (1 + dA(s)) = \exp(A^c(t)) \prod_{0<s\leq t} (1 + \Delta A(s)).$$

The first product is merely notation, but it is motivated by the mathematical definition of the product integral:

$$\phi(A)(t) = \lim_{\max_i |t_i - t_{i-1}| \to 0} \prod_i \{1 + [A(t_i) - A(t_{i-1})]\},$$

where the limit is over partitions $0 = t_0 < t_1 < \cdots t_m = t$ with maximum separation decreasing to zero. We will also use the notation

$$\phi(A)(s,t] = \prod_{s<u\leq t} (1 + dA(u)) \equiv \frac{\phi(A)(t)}{\phi(A)(s)},$$

for all $0 \leq s < t$. The two terms on the left are defined by the far right term. Three alternative definitions of the product integral, as solutions of two different Volterra integral equations and as a "Peano series," are given in Exercise 12.3.2.

The following lemma verifies that product integration is Hadamard differentiable:

LEMMA 12.5 *For fixed constants $0 < b, M < \infty$, the product integral map $\phi : BV_M[0,b] \subset D[0,b] \mapsto D[0,b]$ is Hadamard differentiable with derivative*

$$\phi'_A(\alpha)(t) = \int_{(0,t]} \phi(A)(0,u)\phi(A)(u,t]d\alpha(u).$$

*When $\alpha \in D[0,b]$ has unbounded variation, the above quantity is well-defined by integration by parts.*

We give the proof later on in this section, after first discussing an important statistical application. From the discussion of the Nelson-Aalen estimator $\hat{\Lambda}_n$ in Section 12.2.2, it is not hard to verify that in the right-censored survival analysis setting $S(t) = \phi(-\Lambda)(t)$, where $\phi$ is the product integration map. Moreover, it is easily verified that the Kaplan-Meier estimator $\hat{S}_n$ discussed in Sections 2.2.5 and 4.3 satisfies $\hat{S}_n(t) = \phi(-\hat{\Lambda}_n)(t)$.

We can now use Lemma 12.5 to derive the asymptotic limiting distribution of $\sqrt{n}(\hat{S}_n - S)$. As in Section 12.2.2, we will restrict our time domain to $[0,\tau]$, where $P(X > \tau) > 0$. Under these circumstances, there exists an $M < \infty$, such that $\Lambda(\tau) < M$ and $\hat{\Lambda}_n(\tau) < M$ with probability tending to 1 as $n \to \infty$. Now Lemma 12.5, combined with (12.4) and the discussion immediately following, yields

$$\begin{aligned}\sqrt{n}(\hat{S}_n - S) &\rightsquigarrow -\int_{(0,(\cdot)]} \phi(-\Lambda)(0,u)\phi(-\Lambda)(u,t]\frac{d\mathbb{M}}{Y_0} \\ &= -S(t)\int_{(0,(\cdot)]} \frac{d\mathbb{M}}{(1-\Delta\Lambda)Y_0},\end{aligned}$$

where $\mathbb{M}$ is a Gaussian martingale with independent increments and covariance $\int_{(0,s\wedge t]}(1-\Delta\Lambda)d\Lambda/Y_0$. Thus $\sqrt{n}(\hat{S}_n - S)/S$ is asymptotically time-transformed Brownian motion $\mathbb{W}(w(t))$, where $\mathbb{W}$ is standard Brownian

motion on $[0,\infty)$ and where $w(t) \equiv \int_{(0,t]}[(1-\Delta\Lambda)Y_0]^{-1}d\Lambda$. Along the lines discussed in the Nelson-Aalen example of Section 12.2.2, the form of the limiting distribution can be used to obtain asymptotically exact simultaneous confidence bands for $S$. The delta method bootstrap, Theorem 12.1, can also be used for inference on $S$.

Before giving the proof of Lemma 12.5, we present the following lemma which we will need and which includes the important *Duhamel equation* for the difference between two product integrals:

LEMMA 12.6 *For $A, B \in D(0,b]$, we have for all $0 \le s < t \le b$ the following, where $M$ is the sum of the total variation of $A$ and $B$:*

*(i) (the Duhamel equation)*

$$(\phi(B) - \phi(A))(s,t] = \int_{(s,t]} \phi(A)(0,u)\phi(B)(u,t](B-A)(du).$$

*(ii)* $\|\phi(A) - \phi(B)\|_{(s,t]} \le e^M(1+M)^2\|A-B\|_{(s,t]}$.

**Proof.** For any $u \in (s,t]$, consider the function $C_u \in D(s,t]$ defined as

$$C_u(x) = \begin{cases} A(x) - A(s), \text{ for } s \le x < u, \\ A(u-) - A(s), \text{ for } x = u, \\ A(u-) - A(s) + B(x) - B(u), \text{ for } u < x \le t. \end{cases}$$

Using the Peano series expansion of Exercise 12.3.2, Part (b), we obtain:

$$\begin{aligned} \phi(A)(s,u)\phi(B)(u,t] &= \phi(C_u)(s,t] = 1 \\ &+ \sum_{m,n\ge 0:\, m+n\ge 1} \int_{s<t_1<\cdots<t_m<u<t_{m+1}<\cdots<t_{m+n}\le t} A(dt_1)\cdots A(dt_m) \\ &\times B(dt_{m+1})\cdots B(dt_{m+n}). \end{aligned}$$

Thus

$$\begin{aligned} &\int_{(s,t]} \phi(A)(s,u)\phi(B)(u,t](B-A)(du) \\ &= \sum_{n\ge 1} \int_{s<x_1<\cdots<x_n\le t} \left[1 + \sum_{m\ge 1} \int_{s<t_1<\cdots<t_m<x_1} A(dt_1)\cdots A(dt_m)\right] \\ &\qquad \times B(dx_1)\cdots B(dx_n) \\ &\quad - \sum_{n\ge 1} \int_{s<t_1<\cdots<t_n\le t} \left[1 + \sum_{m\ge 1} \int_{t_n<x_1<\cdots<x_m\le t} B(dx_1)\cdots B(dx_n)\right] \\ &\qquad \times A(dt_1)\cdots A(dt_n) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{n\geq 1} \int_{s<x_1<\cdots<x_n\leq t} B(dx_1)\cdots B(dx_n) \\
&\quad - \sum_{n\geq 1} \int_{s<t_1<\cdots t_n\leq t} A(dt_1)\cdots A(dt_n) \\
&= \phi(B)(s,t] - \phi(A)(s,t].
\end{aligned}
$$

This proves Part (i).

For Part (ii), we need to derive an integration by parts formula for the Duhamel equation. Define $G = B - A$ and $H(u) \equiv \int_0^u \phi(B)(v,t]G(dv)$. Now integration by parts gives us

$$
\begin{aligned}
&\int_{(s,t]} \phi(A)(0,u)\phi(B)(u,t]G(du) \\
&\quad = \phi(A)(t)H(t) - \phi(A)(s)H(s) - \int_{(s,t]} H(u)\phi(A)(du).
\end{aligned} \tag{12.6}
$$

From the backwards integral equation (Part (c) of Exercise 12.3.2), we know that $\phi(B)(dv,t] = -\phi(B)(v,t]B(dv)$, and thus, by integration by parts, we obtain

$$
H(u) = G(u)\phi(B)(u,t] + \int_{(0,u]} G(v-)\phi(B)(v,t]B(dv).
$$

Combining this with (12.6) and the fact that $\phi(A)(du) = \phi(A)(u-)A(du)$, we get

$$
\begin{aligned}
&\int_{(s,t]} \phi(A)(0,u)\phi(B)(u,t]G(du) \\
&\quad = \phi(A)(t)\int_{(0,t]} G(u-)\phi(B)(u,t]B(du) \\
&\qquad -\phi(A)(s)\phi(B)(s,t]G(s) \\
&\qquad -\phi(A)(s)\int_{(0,s]} G(u-)\phi(B)(u,t]B(du) \\
&\qquad -\int_{(s,t]} G(u)\phi(B)(u,t]\phi(A)(u-)A(du) \\
&\qquad -\int_{(s,t]}\int_{(0,u]} G(v-)\phi(B)(v,t]B(dv)\phi(A)(u-)A(du).
\end{aligned} \tag{12.7}
$$

From Exercise 12.3.3, we know that $\phi(A)$ and $\phi(B)$ are bounded by the exponentiation of the respective total variations of $A$ and $B$. Now the desired result follows.□

**Proof of Lemma 12.5.** Set $A_n = A + t_n\alpha_n$ for a sequence $\alpha_n \to \alpha$ with the total variation of both $A$ and $A_n$ bounded by $M$. In view of the Duhamel equation (Part (i) of Lemma 12.6 above), it suffices to show that

$$\int_{(0,t]} \phi(A)(0,u)\phi(A_n)(u,t]d\alpha_n(u) \to \int_{(0,t]} \phi(A)(0,u)\phi(A)(u,t]d\alpha(u),$$

uniformly in $0 \le t \le b$. Fix $\epsilon > 0$. Since $\alpha \in D[0,b]$, there exists a function $\tilde{\alpha}$ with total variation $V < \infty$ such that $\|\alpha - \tilde{\alpha}\|_\infty \le \epsilon$.

Now recall that the derivation of the integration by parts formula (12.7) from the proof of Lemma 12.6 did not depend on the definition of $G$, other than the necessity of $G$ being right-continuous. If we replace $G$ with $\alpha - \tilde{\alpha}$, we obtain from (12.7) that

$$\begin{aligned}\left\| \int_{(0,(\cdot)]} \phi(A)(0,u)\phi(A_n)(u,t]d(\alpha_n - \tilde{\alpha})(u) \right\|_\infty & \\ \le \quad & e^{2M}(1+2M)^2\|\alpha_n - \tilde{\alpha}\|_\infty \\ \to \quad & e^{2M}(1+2M)^2\epsilon,\end{aligned}$$

as $n \to \infty$, since $\|\alpha_n - \alpha\|_\infty \to 0$. Moreover,

$$\begin{aligned}\left\| \int_{(0,(\cdot)]} \phi(A)(0,u) \left[\phi(A_n) - \phi(A)\right](u,t]d\tilde{\alpha}(u) \right\|_\infty & \\ \le \quad & \|\phi(A_n) - \phi(A)\|_\infty \|\phi(A)\|_\infty V \\ \to \quad & 0,\end{aligned}$$

as $n \to \infty$. Now using again the integration by parts formula (12.7), but with $G = \tilde{\alpha} - \alpha$, we obtain

$$\left\| \int_{0,(\cdot)]} \phi(A)(0,u)\phi(A)(u,t]d(\tilde{\alpha} - \alpha)(u) \right\|_\infty \le e^{2M}(1+2M)^2\epsilon.$$

Thus the desired result follows since $\epsilon$ was arbitrary.□

### *12.2.4 Inversion*

Recall the derivation given in the paragraphs following Theorem 2.8 of the Hadamard derivative of the inverse of a distribution function $F$. Note that this derivation did not depend on $F$ being a distribution function per se. In fact, the derivation will carry through unchanged if we replace the distribution function $F$ with any nondecreasing, cadlag function $A$ satisfying mild regularity conditions. For a non-decreasing function $B \in D(-\infty,\infty)$, define the left-continuous inverse $r \mapsto B^{-1}(r) \equiv \inf\{x : B(x) \ge r\}$. We will hereafter use the notation $\tilde{D}[u,v]$ to denote all left-continuous functions with right-hand limits (caglad) on $[u,v]$ and $D_1[u,v]$ to denote the restriction of all non-decreasing functions in $D(-\infty,\infty)$ to the interval $[u,v]$. Here is a precise statement of the general Hadamard differentiation result for non-decreasing functions:

LEMMA 12.7 *Let $-\infty < p \leq q < \infty$, and let the non-decreasing function $A \in D(-\infty, \infty)$ be continuously differentiable on the interval $[u, v] \equiv [A^{-1}(p) - \epsilon, A^{-1}(q) + \epsilon]$, for some $\epsilon > 0$, with derivative $A'$ being strictly positive and bounded over $[u, v]$. Then the inverse map $B \mapsto B^{-1}$ as a map $D_1[u, v] \subset D[u, v] \mapsto \tilde{D}[p, q]$ is Hadamard differentiable at $A$ tangentially to $C[u, v]$, with derivative $\alpha \mapsto -(\alpha/A') \circ A^{-1}$.*

We will give the proof of Lemma 12.7 at the end of this section. We now restrict ourselves to the setting where $A$ is a distribution function which we will now denote by $F$. The following lemma provides two results similar to Lemma 12.7 but which utilize knowledge about the support of the distribution function $F$. Let $D_2[u, v]$ be the subset of distribution functions in $D_1[u, v]$ with support only on $[u, \infty)$, and let $D_3[u, v]$ be the subset of distribution functions in $D_2[u, v]$ which have support only on $[u, v]$.

LEMMA 12.8 *Let $F$ be a distribution function. We have the following:*

*(i) Let $F \in D_2[u, \infty)$, for finite $u$, and let $q \in (0, 1)$. Assume $F$ is continuously differentiable on the interval $[u, v] = [u, F^{-1}(q) + \epsilon]$, for some $\epsilon > 0$, with derivative $f$ being strictly positive and bounded over $[u, v]$. Then the inverse map $G \mapsto G^{-1}$ as a map $D_2[u, v] \subset D[u, v] \mapsto \tilde{D}(0, q]$ is Hadamard differentiable at $F$ tangentially to $C[u, v]$.*

*(ii) Let $F \in D_3[u, v]$, for $[u, v]$ compact, and assume that $F$ is continuously differentiable on $[u, v]$ with derivative $f$ strictly positive and bounded over $[u, v]$. Then the inverse map $G \mapsto G^{-1}$ as a map $D_3[u, v] \subset D[u, v] \mapsto \tilde{D}(0, 1)$ is Hadamard differentiable at $F$ tangentially to $C[u, v]$.*

*In either case, the derivative is the map $\alpha \mapsto -(\alpha/f) \circ F^{-1}$.*

Before giving the proof of the above two lemmas, we will discuss some important statistical applications. As discussed in Section 2.2.4, an important application of these results is to estimation and inference for the quantile function $p \mapsto F^{-1}(p)$ based on the usual empirical distribution function for i.i.d. data. Lemma 12.8 is useful when some information is available on the support of $F$, since it allows the range of $p$ to extend as far as possible. These results are applicable to other estimators of the distribution function $F$ besides the usual empirical distribution, provided the standardized estimators converge to a tight limiting process over the necessary intervals. Several examples of such estimators include the Kaplan-Meier estimator, the self-consistent estimator of Chang (1990) for doubly-censored data, and certain estimators from dependent data as mentioned in Kosorok (1999).

We now apply Lemma 12.8 to the construction of quantile processes based on the Kaplan-Meier estimator discussed in Section 12.2.3 above. Since it is known that the support of a survival function is on $[0, \infty)$, we can utilize Part (i) of this lemma. Define the Kaplan-Meier quantile

process $\{\hat{\xi}(p) \equiv \hat{F}_n^{-1}(p), 0 < p \leq q\}$, where $\hat{F}_n = 1 - \hat{S}_n$, $\hat{S}_n$ is the Kaplan-Meier estimator, and where $0 < q < F(\tau)$ for $\tau$ as defined in the previous section. Assume that $F$ is continuously differentiable on $[0, \tau]$ with density $f$ bounded below by zero and finite. Combining the results of the previous section with Part (i) of Lemma 12.8 and Theorem 2.8, we obtain

$$\sqrt{n}(\hat{\xi} - \xi)(\cdot) \rightsquigarrow \frac{S(\xi(\cdot))}{f(\xi(\cdot))} \int_{(0,\xi(\cdot)]} \frac{d\mathbb{M}}{(1 - \Delta\Lambda)Y_0},$$

in $\tilde{D}(0, q]$, where $\xi(p) \equiv \xi_p$ and $\mathbb{M}$ is the Gaussian martingale described in the previous section. Thus $\sqrt{n}(\hat{\xi} - \xi)f(\xi)/S(\xi)$ is asymptotically time-transformed Brownian motion with time-transform $w(\xi)$, where $w$ is as defined in the previous section, over the interval $(0, q]$. As described in Kosorok (1999), one can construct kernel estimators for $f$—which can be shown to be uniformly consistent—to facilitate inference. An alternative approach is the bootstrap which can be shown to be valid in this setting based on Theorem 12.1.

**Proof of Lemma 12.7.** The arguments are essentially identical to those used in the paragraphs following Theorem 2.8, except that the distribution function $F$ is replaced by a more general, non-decreasing function $A$.□

**Proof of Lemma 12.8.** To prove Part (i), let $\alpha_n \to \alpha$ uniformly in $D[u, v]$ and $t_n \to 0$, where $\alpha$ is continuous and $F + t_n\alpha_n$ is contained in $D_2[u, v]$ for all $n \geq 1$. Abbreviate $F^{-1}(p)$ and $(F + t_n\alpha_n)^{-1}(p)$ to $\xi_p$ and $\xi_{pn}$, respectively. Since $F$ and $F + t_n\alpha_n$ have domains $(u, \infty)$ (the lower bound by assumption), we have that $\xi_p, \xi_{pn} > u$ for all $0 < p \leq q$. Moreover, $\xi_p, \xi_{pn} \leq v$ for all $n$ large enough. Thus the numbers $\epsilon_{pn} \equiv t_n^2 \wedge (\xi_{pn} - u)$ are positive for all $0 < p \leq q$, for all $n$ large enough. Hence, by definition, we have for all $p \in (0, q]$ that

$$(F + t_n\alpha_n)(\xi_{pn} - \epsilon_{pn}) \leq p \leq (F + t_n\alpha_n)(\xi_{pn}), \tag{12.8}$$

for all sufficiently large $n$.

By the smoothness of $F$, we have $F(\xi_p) = p$ and $F(\xi_{pn} - \epsilon_{pn}) = F(\xi_{pn}) + O(\epsilon_{pn})$, uniformly over $p \in (0, q]$. Thus from (12.8) we obtain

$$-t_n\alpha(\xi_{pn}) + o(t_n) \leq F(\xi_{pn}) - F(\xi_p) \leq -t_n\alpha(\xi_{pn} - \epsilon_{pn}) + o(t_n), \tag{12.9}$$

where the $o(t_n)$ terms are uniform over $0 < p \leq q$. Both the far left and far right sides are $O(t_n)$, while the middle term is bounded above and below by constants times $|\xi_{pn} - \xi_p|$, for all $0 < p \leq q$. Hence $|\xi_{pn} - \xi_p| = O(t_n)$, uniformly over $0 < p \leq q$. The Part (i) result now follows from (12.9), since $F(\xi_{pn}) - F(\xi_p) = f(\xi_p)(\xi_{pn} - \xi_p) + E_n$, where $E_n = o(\sup_{0<p\leq q} |\xi_{pn} - \xi_p|)$ by the uniform differentiability of $F$ over $(u, v]$.

Note that Part (ii) of this lemma is precisely Part (ii) of Lemma 3.9.23 of VW, and the details of the proof (which are quite similar to the proof of Part (i)) can be found therein.□

### *12.2.5 Other Mappings*

We now mention briefly a few additional interesting examples. The first example is the *copula map*. For a bivariate distribution function $H$, with marginals $F_H(x) \equiv H(x, \infty)$ and $G_H(y) \equiv H(\infty, y)$, the copula map is the map $\phi$ from bivariate distributions on $\mathbb{R}^2$ to bivariate distributions on $[0,1]^2$ defined as follows:

$$H \mapsto \phi(H)(u,v) = H(F_H^{-1}(u), G_H^{-1}(v)), \quad (u,v) \in [0,1]^2,$$

where the inverse functions are the left-continuous quantile functions defined in the previous section. Section 3.9.4.4 of VW verifies that this map is Hadamard differentiable in a manner which permits developing inferential procedures for the copula function based on i.i.d. bivariate data.

The second example is multivariate trimming. Let $P$ be a given probability distribution on $\mathbb{R}^d$, fix $\alpha \in (0, 1/2]$, and define $\mathcal{H}$ to be the collection of all closed half-spaces in $\mathbb{R}^d$. The set $K_P \equiv \cap\{H \in \mathcal{H} : P(H) \geq 1 - \alpha\}$ can be easily shown to be compact and convex (see Exercise 12.3.4). The *$\alpha$-trimmed mean* is the quantity

$$T(P) \equiv \frac{1}{\lambda(K_P)} \int_{K_P} x d\lambda(x),$$

where $\lambda$ is the Lebesgue measure on $\mathbb{R}^d$. Using non-trivial arguments, Section 3.9.4.6 of VW shows how $P \mapsto T(P)$ can be formulated as a Hadamard differentiable functional of $P$ and how this formulation can be applied to develop inference for $T(P)$ based on i.i.d. data from $P$.

There are many other important examples in statistics, some of which we will explore later on in this book, including a delta method formulation of Z-estimator theory which we will describe in the next chapter (Chapter 13) and several other examples in the case studies of Chapter 15.

## 12.3 Exercises

**12.3.1.** In the Wilcoxon statistic example of Section 12.2.2, verify explicitly that every hypothesis of Theorem 2.8 is satisfied.

**12.3.2.** Show that the product integral of $A$, $\phi(A)(s,t]$, is equivalent to the following:

(a) The unique solution $B$ of the following Volterra integral equation:

$$B(s,t] = 1 + \int_{(s,t]} B(s,u) A(du).$$

(b) The following **Peano series** representation:

$$\phi(A)(s,t] = 1 + \sum_{m=1}^{\infty} \int_{s<t_1<\cdots<t_m\le t} A(dt_1)\cdots A(dt_m),$$

where the signed-measure interpretation of $A$ is being used. Hint: Use the uniqueness from Part (a).

(c) The unique solution $B$ of the "backward" Volterra integral equation:

$$B(s,t] = 1 + \int_{(s,t]} B(u,t]A(du).$$

Hint: Start at $t$ and go backwards in time to $s$.

**12.3.3.** Let $\phi$ be the product integral map of Section 12.2.3. Show that if the total variation of $A$ over the interval $(s,t]$ is $M$, then $|\phi(A)(s,t]| \le e^M$. Hint: Recall that $\log(1+x) \le x$ for all $x > 0$.

**12.3.4.** Show that the set $K_P \subset \mathbb{R}^d$ defined in Section 12.2.5 is compact and convex.

## 12.4 Notes

Much of the material of this chapter is inspired by Chapter 3.9 of VW, although there is some new material and the method of presentation is different. Section 12.1 contains results from Sections 3.9.1 and 3.9.3 of VW, although our results are specialized to Banach spaces (rather than the more general topological vector spaces). The examples of Sections 12.2.1 through 12.2.4 are modified versions of the examples of Sections 3.9.4.3, 3.9.4.1, 3.9.4.5 and 3.9.4.2, respectively, of VW. The order has be changed to emphasize a natural progression leading up to quantile inference based on the Kaplan-Meier estimator. Lemma 12.2 is a generalization of Lemma 3.9.25 of VW, while Lemmas 12.3 and 12.5 correspond to Lemmas 3.9.17 and 3.9.30 of VW. Lemma 12.7 is a generalization of Part (i) of Lemma 3.9.23 of VW, while Part (ii) of Lemma 12.8 corresponds to Part (ii) of Lemma 3.9.23 of VW. Exercise 12.3.2 is based on Exercises 3.9.5 and 3.9.6 of VW.

# 13
# Z-Estimators

Recall from Section 2.2.5 that Z-estimators are approximate zeros of data-dependent functions. These data-dependent functions, denoted $\Psi_n$, are maps between a possibly infinite dimensional normed parameter space $\Theta$ and a normed space $\mathbb{L}$, where the respective norms are $\|\cdot\|$ and $\|\cdot\|_{\mathbb{L}}$. The $\Psi_n$ are frequently called estimating equations. A quantity $\hat{\theta}_n \in \Theta$ is a Z-estimator if $\|\Psi_n(\hat{\theta}_n)\|_{\mathbb{L}} \xrightarrow{\mathrm{P}} 0$. In this chapter, we extend and prove the results of Section 2.2.5. As part of this, we extend the Z-estimator master theorem, Theorem 10.16, to the infinite dimensional case, although we divide the result into two parts, one for consistency and one for weak convergence.

We first discuss consistency and present a Z-estimator master theorem for consistency. We then discuss weak convergence and examine closely the special case of Z-estimators which are empirical measures of Donsker classes. We then use this structure to develop a Z-estimator master theorem for weak convergence. Both master theorems, the one for consistency and the one for weak convergence, will include results for the bootstrap. Finally, we demonstrate how Z-estimators can be viewed as Hadamard differentiable functionals of the involved estimating equations and how this structure enables use of a modified delta method to obtain very general results for Z-estimators. Recall from Section 2.2.5 that the Kaplan-Meier estimator is an important and instructive example of a Z-estimator. A more sophisticated example, which will be presented later in the case studies of Chapter 15, is the nonparametric maximum likelihood estimator for the proportional odds survival model.

## 13.1 Consistency

The main consistency result we have already presented in Theorem 2.10 of Section 2.2.5, and the proof of this theorem was given as an exercise (Exercise 2.4.2). We will now extend this result to the bootstrapped Z-estimator. First, we restate the identifiability condition of Theorem 2.10: The map $\Psi : \Theta \mapsto \mathbb{L}$ is identifiable at $\theta_0 \in \Theta$ if

$$\|\Psi(\theta_n)\|_{\mathbb{L}} \to 0 \text{ implies } \|\theta_n - \theta_0\| \to 0 \text{ for any } \{\theta_n\} \in \Theta. \tag{13.1}$$

Note that there are alternative identifiability conditions that will also work, including the stronger condition that both $\Psi(\theta_0) = 0$ and $\Psi : \Theta \mapsto \mathbb{L}$ be one-to-one. Nevertheless, Condition (13.1) seems to be the most efficient for our purposes.

In what follows, we will use the bootstrap-weighted empirical process $\mathbb{P}_n^\circ$ to denote either the nonparametric bootstrapped empirical process (with multinomial weights) or the multiplier bootstrapped empirical process defined by $f \mapsto \mathbb{P}_n^\circ f = n^{-1}\sum_{i=1}^n (\xi_i/\bar{\xi}) f(X_i)$, where $\xi_1, \ldots, \xi_n$ are i.i.d. positive weights with $0 < \mu = \mathrm{E}\xi_1 < \infty$ and $\bar{\xi} = n^{-1}\sum_{i=1}^n \xi_i$. Note that this is a special case of the weighted bootstrap introduced in Theorem 10.13 but with the addition of $\bar{\xi}$ in the denominator. We leave it as an exercise (Exercise 13.4.1) to verify that the conclusions of Theorem 10.13 are not affected by this addition. Let $\mathcal{X}_n \equiv \{X_1, \ldots, X_n\}$ as given in Theorem 10.13. The following is the main result of this section:

THEOREM 13.1 *(Master Z-estimator theorem for consistency) Let $\theta \mapsto \Psi(\theta) = P\psi_\theta$, $\theta \mapsto \Psi_n(\theta) = \mathbb{P}_n\psi_\theta$ and $\theta \mapsto \Psi_n^\circ(\theta) = \mathbb{P}_n^\circ\psi_\theta$, where $\Psi$ satisfies (13.1) and the class $\{\psi_\theta : \theta \in \Theta\}$ is P-Glivenko-Cantelli. Then, provided $\|\Psi_n(\hat{\theta}_n)\|_{\mathbb{L}} = o_P(1)$ and*

$$\mathrm{P}\left(\|\Psi_n^\circ(\hat{\theta}_n^\circ)\|_{\mathbb{L}} > \eta \middle| \mathcal{X}_n\right) = o_P(1) \text{ for every } \eta > 0, \tag{13.2}$$

*we have both $\|\hat{\theta}_n - \theta_0\| = o_P(1)$ and $\mathrm{P}\left(\|\hat{\theta}_n^\circ - \theta_0\| > \eta \middle| \mathcal{X}_n\right) = o_P(1)$ for every $\eta > 0$.*

Note in (13.2) the absence of an outer probability on the left side. This is because, as argued in Section 2.2.3, a Lipschitz continuous map of either of these bootstrapped empirical processes is measurable with respect to the random weights conditional on the data.

**Proof of Theorem 13.1.** The result that $\|\hat{\theta}_n - \theta_0\| = o_P(1)$ is a conclusion from Theorem 2.10. For the conditional bootstrap result, (13.2) implies that for some sequence $\eta_n \downarrow 0$, $P\left(\|\Psi(\hat{\theta}_n^\circ)\|_{\mathbb{L}} > \eta_n \middle| \mathcal{X}_n\right) = o_P(1)$, since

$$P\left(\sup_{\theta\in\Theta}\|\Psi_n^\circ(\theta) - \Psi(\theta)\| > \eta \middle| \mathcal{X}_n\right) = o_P(1)$$

for all $\eta > 0$, by Theorems 10.13 and 10.15. Thus, for any $\epsilon > 0$,

$$\begin{aligned} \mathrm{P}\left( \|\hat{\theta}_n^\circ - \theta_0\| > \epsilon \middle| \mathcal{X}_n \right) &\leq \mathrm{P}\left( \|\hat{\theta}_n^\circ - \theta_0\| > \epsilon, \|\Psi(\hat{\theta}_n^\circ)\|_{\mathbb{L}} \leq \eta_n \middle| \mathcal{X}_n \right) \\ &\quad + P\left( \|\Psi(\hat{\theta}_n^\circ)\|_{\mathbb{L}} > \eta_n \middle| \mathcal{X}_n \right) \\ &\xrightarrow{\mathrm{P}} 0, \end{aligned}$$

since the identifiability Condition (13.1) implies that for all $\delta > 0$ there exists an $\eta > 0$ such that $\|\Psi(\theta)\|_{\mathbb{L}} < \eta$ implies $\|\theta - \theta_0\| < \delta$. Hence it is impossible for there to exist any $\theta \in \Theta$ such that both $\|\theta - \theta_0\| > \epsilon$ and $\|\Psi(\theta)\|_{\mathbb{L}} < \eta_n$ for all $n \geq 1$. The conclusion now follows since $\epsilon$ was arbitrary.□

Note that we might have worked toward obtaining outer almost sure results since we are making a strong Glivenko-Cantelli assumption for the class of functions involved. However, we only need convergence in probability for statistical applications. Notice also that we only assumed $\|\Psi_n^\circ(\hat{\theta}_n^\circ)\|$ goes to zero conditionally rather than unconditionally as done in Theorem 10.16. However, it seems to be easier to check the conditional version in practice. Moreover, the unconditional version is actually stronger than the conditional version, since

$$\mathrm{E}^* \mathrm{P}\left( \|\Psi_n^\circ(\hat{\theta}_n^\circ)\|_{\mathbb{L}} > \eta \middle| \mathcal{X}_n \right) \leq \mathrm{P}^*\left( \|\Psi_n^\circ(\hat{\theta}_n^\circ)\|_{\mathbb{L}} > \eta \right)$$

by the version of Fubini's theorem given as Theorem 6.14. It is unclear how to extend this argument to the outer almost sure setting. This is another reason for restricting our attention to the convergence in probability results. Nevertheless, we still need the strong Glivenko-Cantelli assumption since this enables the use of Theorems 10.13 and 10.15.

While the above approach will be helpful for some Z-estimators, many Z-estimators are complex enough to require individually Tailored approaches to establishing consistency. In Section 13.2.3 below, we will revisit the Kaplan-Meier estimator example of Section 2.2.5 to which we can apply a generalization of Theorem 13.1, Theorem 13.4, which includes weak convergence. In contrast, the proportional odds model for right-censored survival data, which will be presented in Chapter 15, requires a more individualized approach to establishing consistency.

## 13.2 Weak Convergence

In this section, we first provide general results for Z-estimators which may not be based on i.i.d. data. We then present sufficient conditions for the i.i.d. case when the estimating equation is an empirical measure ranging over a Donsker class. Finally, we give a master theorem for Z-estimators based on i.i.d. data which includes bootstrap validity.

### 13.2.1 The General Setting

We now prove Theorem 2.11 on Page 26 and give a method of weakening the differentiability requirement for $\Psi$. An important thing to note is that no assumptions about the data being i.i.d. are required. The proof follows closely the proof of Theorem 3.3.1 given in VW.

**Proof of Theorem 2.11 (Page 26).** By the definitions of $\hat{\theta}_n$ and $\theta_0$,

$$\begin{aligned} \sqrt{n}\left(\Psi(\hat{\theta}_n) - \Psi(\theta_0)\right) &= -\sqrt{n}\left(\Psi_n(\hat{\theta}_n) - \Psi(\hat{\theta}_n)\right) + o_P(1) \\ &= -\sqrt{n}(\Psi_n - \Psi)(\theta_0) + o_P(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|), \end{aligned} \tag{13.3}$$

by Assumption (2.12). Note the error terms throughout this theorem are with respect to the norms of the spaces, e.g. $\Theta$ or $\mathbb{L}$, involved. Since $\dot{\Psi}_{\theta_0}$ is continuously invertible, we have by Part (i) of Lemma 6.16 that there exists a constant $c > 0$ such that $\|\dot{\Psi}_{\theta_0}(\theta - \theta_0)\| \geq c\|\theta - \theta_0\|$ for all $\theta$ and $\theta_0$ in $\overline{\text{lin}}\,\Theta$. Combining this with the differentiability of $\Psi$ yields $\|\Psi(\theta) - \Psi(\theta_0)\| \geq c\|\theta - \theta_0\| + o(\|\theta - \theta_0\|)$. Combining this with (13.3), we obtain

$$\sqrt{n}\|\hat{\theta}_n - \theta_0\|(c + o_P(1)) \leq O_P(1) + o_P(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|).$$

We now have that $\hat{\theta}_n$ is $\sqrt{n}$-consistent for $\theta_0$ with respect to $\|\cdot\|$. By the differentiability of $\Psi$, the left side of (13.3) can be replaced by $\sqrt{n}\dot{\Psi}_{\theta_0}(\hat{\theta}_n - \theta_0) + o_P(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|)$. This last error term is now $o_P(1)$ as also is the error term on the right side of (13.3). Now the result (2.13) follows. Next the continuity of $\dot{\Psi}_{\theta_0}^{-1}$ and the continuous mapping theorem yield $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow -\dot{\Psi}_{\theta_0}^{-1}(Z)$ as desired.□

The following lemma allows us to weaken the Fréchet differentiability requirement to Hadamard differentiability when it is also known that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically tight:

LEMMA 13.2 *Assume the conditions of Theorem 2.11 except that consistency of $\hat{\theta}_n$ is strengthened to asymptotic tightness of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ and the Fréchet differentiability of $\Psi$ is weakened to Hadamard differentiability at $\theta_0$. Then the results of Theorem 2.11 still hold.*

**Proof.** The asymptotic tightness of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ enables Expression (13.3) to imply $\sqrt{n}\left(\Psi(\hat{\theta}_n) - \Psi(\theta_0)\right) = -\sqrt{n}(\Psi_n - \Psi)(\theta_0) + o_P(1)$. The Hadamard differentiability of $\Psi$ yields $\sqrt{n}\left(\Psi(\hat{\theta}_n) - \Psi(\theta_0)\right) = \sqrt{n}\dot{\Psi}_{\theta_0}(\hat{\theta}_n - \theta_0) + o_P(1)$. Combining, we now have $\sqrt{n}\dot{\Psi}_{\theta_0}(\hat{\theta}_n - \theta_0) = -\sqrt{n}(\Psi_n - \Psi)(\theta_0) + o_P(1)$, and all of the results of the theorem follow.□

### 13.2.2 Using Donsker Classes

We now consider the special case where the data involved are i.i.d., i.e., $\Psi_n(\theta)(h) = \mathbb{P}_n\psi_{\theta,h}$ and $\Psi(\theta)(h) = P\psi_{\theta,h}$, for measurable functions $\psi_{\theta,h}$,

where $h$ ranges over an index set $\mathcal{H}$. The following lemma gives us reasonably verifiable sufficient conditions for (2.12) to hold:

LEMMA 13.3 *Suppose the class of functions*

$$\{\psi_{\theta,h} - \psi_{\theta_0,h} : \|\theta - \theta_0\| < \delta, h \in \mathcal{H}\} \tag{13.4}$$

*is $P$-Donsker for some $\delta > 0$ and*

$$\sup_{h\in\mathcal{H}} P\left(\psi_{\theta,h} - \psi_{\theta_0,h}\right)^2 \quad \to \quad 0, \quad \text{as } \theta \to \theta_0. \tag{13.5}$$

*Then if* $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$, $\sup_{h\in\mathcal{H}} \left|\mathbb{G}_n\psi_{\hat{\theta}_n,h} - \mathbb{G}_n\psi_{\theta_0,h}\right| = o_P(1)$.

Before giving the proof of this lemma, we make the somewhat trivial observation that the conclusion of this lemma implies (2.12).

**Proof of Lemma 13.3.** Let $\Theta_\delta \equiv \{\theta : \|\theta - \theta_0\| < \delta\}$ and define the extraction function $f : \ell^\infty(\Theta_\delta \times \mathcal{H}) \times \Theta_\delta \mapsto \ell^\infty(\mathcal{H})$ as $f(z,\theta)(h) \equiv z(\theta,h)$, where $z \in \ell^\infty(\Theta_\delta \times \mathcal{H})$. Note that $f$ is continuous at every point $(z,\theta_1)$ such that $\sup_{h\in\mathcal{H}} |z(\theta,h) - z(\theta_1,h)| \to 0$ as $\theta \to \theta_1$. Define the stochastic process $Z_n(\theta,h) \equiv \mathbb{G}_n(\psi_{\theta,h} - \psi_{\theta_0,h})$ indexed by $\Theta_\delta \times \mathcal{H}$. As assumed, the process $Z_n$ converges weakly in $\ell^\infty(\Theta_\delta \times \mathcal{H})$ to a tight Gaussian process $Z_0$ with continuous sample paths with respect to the metric $\rho$ defined by $\rho^2((\theta_1,h_1),(\theta_2,h_2)) = P(\psi_{\theta_1,h_1} - \psi_{\theta_0,h_1} - \psi_{\theta_2,h_2} + \psi_{\theta_0,h_2})^2$. Since, $\sup_{h\in\mathcal{H}} \rho((\theta,h),(\theta_0,h)) \to 0$ by assumption, we have that $f$ is continuous at almost all sample paths of $Z_0$. By Slutsky's theorem (Theorem 7.15), $(Z_n, \hat{\theta}_n) \rightsquigarrow (Z_0,\theta_0)$. The continuous mapping theorem (Theorem 7.7) now implies that $Z_n(\hat{\theta}_n) = f(Z_n,\hat{\theta}_n) \rightsquigarrow f(Z_0,\theta_0) = 0$.□

If, in addition to the assumptions of Lemma 13.3, we are willing to assume

$$\{\psi_{\theta_0,h} : h \in \mathcal{H}\} \tag{13.6}$$

is $P$-Donsker, then $\sqrt{n}(\Psi_n - \Psi)(\theta_0) \rightsquigarrow Z$, and all of the weak convergence assumptions of Theorem 2.11 are satisfied. Alternatively, we could just assume that

$$\mathcal{F}_\delta \quad \equiv \quad \{\psi_{\theta,h} : \|\theta - \theta_0\| < \delta, h \in \mathcal{H}\} \tag{13.7}$$

is $P$-Donsker for some $\delta > 0$, then both (13.4) and (13.6) are $P$-Donsker for some $\delta > 0$. We are now well poised for a Z-estimator master theorem for weak convergence.

### *13.2.3 A Master Theorem and the Bootstrap*

In this section, we augment the results of the previous section to achieve a general Z-estimator master theorem that includes both weak convergence

and validity of the bootstrap. Here we consider the two bootstrapped Z-estimators described in Section 13.1, except that for the multiplier bootstrap we make the additional requirements that $0 < \tau^2 = \mathrm{var}(\xi_1) < \infty$ and $\|\xi_1\|_{2,1} < \infty$. We use $\overset{\mathrm{P}}{\underset{\circ}{\rightsquigarrow}}$ to denote either $\overset{\mathrm{P}}{\underset{\xi}{\rightsquigarrow}}$ or $\overset{\mathrm{P}}{\underset{W}{\rightsquigarrow}}$ depending on which bootstrap is being used, and we let the constant $k_0 = \tau/\mu$ for the multiplier bootstrap and $k_0 = 1$ for the multinomial bootstrap. Here is the main result:

THEOREM 13.4 *Assume $\Psi(\theta_0) = 0$ and the following hold:*

*(A) $\theta \mapsto \Psi(\theta)$ satisfies (13.1);*

*(B) The class $\{\psi_{\theta,h}; \theta \in \Theta, h \in \mathcal{H}\}$ is P-Glivenko-Cantelli;*

*(C) The class $\mathcal{F}_\delta$ in (13.7) is P-Donsker for some $\delta > 0$;*

*(D) Condition (13.5) holds;*

*(E) $\|\Psi_n(\hat{\theta}_n)\|_{\mathbb{L}} = o_P(n^{-1/2})$ and $\mathrm{P}\left(\sqrt{n}\|\Psi_n^\circ(\hat{\theta}_n^\circ)\|_{\mathbb{L}} > \eta \,\middle|\, \mathcal{X}_n\right) = o_P(1)$ for every $\eta > 0$;*

*(F) $\theta \mapsto \Psi(\theta)$ is Fréchet-differentiable at $\theta_0$ with continuously invertible derivative $\dot{\Psi}_{\theta_0}$.*

*Then $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow -\dot{\Psi}_{\theta_0}^{-1} Z$, where $Z \in \ell^\infty(\mathcal{H})$ is the tight, mean zero Gaussian limiting distribution of $\sqrt{n}(\Psi_n - \Psi)(\theta_0)$, and $\sqrt{n}(\hat{\theta}_n^\circ - \hat{\theta}_n) \overset{\mathrm{P}}{\underset{\circ}{\rightsquigarrow}} k_0 Z$.*

Condition (A) is identifiability. Conditions (B) and (C) are consistency and asymptotic normality conditions for the estimating equation. Condition (D) is an asymptotic equicontinuity condition for the estimating equation at $\theta_0$. Condition (E) simply states that the estimators are approximate zeros of the estimating equation, while Condition (F) specifies the smoothness and invertibility requirements of the derivative of $\Psi$. Except for the last half of Condition (E), all of the conditions are requirements for asymptotic normality of $\sqrt{n}(\hat{\theta}_n - \theta_0)$. What is perhaps surprising is how little additional assumptions are needed to obtain bootstrap validity. Only an assurance that the bootstrapped estimator is an approximate zero of the bootstrapped estimating equation is required. Thus bootstrap validity is almost an automatic consequence of asymptotic normality.

Before giving the proof of the theorem, we will present an example. Recall the right-censored Kaplan-Meier estimator example of Section 2.2.5 which was shown to be a Z-estimator with a certain estimating equation $\Psi_n(\theta) = \mathbb{P}_n \psi_\theta(t)$, where

$$\psi_\theta(t) = 1\{U > t\} + (1-\delta)1\{U \le t\}1\{\theta(U) > 0\}\frac{\theta(t)}{\theta(U)} - \theta(t),$$

where the observed data $U, \delta$ is the right-censored survival time and censoring indicator, respectively, and where we have replaced the survival function $S$ with $\theta$ in the notation of Section 2.2.5 to obtain greater consistency with the notation of the current chapter. The limiting estimating function $\Psi(\theta) = P\psi_\theta$ is given in (2.11). It was shown in the paragraphs surrounding (2.11) that both $\Psi(\theta_0) = 0$ and $\Psi(\theta_n) \to 0$ imply $\|\theta_n - \theta_0\|_\infty \to 0$, and thus the opening line and Condition (A) of the theorem are satisfied. Exercise 2.4.3 verifies that $\Psi(\theta)$ is Fréchet differentiable with derivative $\dot{\Psi}_{\theta_0}$ defined in (2.16). Exercise 2.4.4 verifies that $\dot{\Psi}_{\theta_0}$ is continuously invertible with inverse $\dot{\Psi}_{\theta_0}^{-1}$ given explicitly in (2.17). Thus Condition (F) of the theorem is also established.

In the paragraphs in Section 2.2.5 after the presentation of Theorem 2.1, the class $\{\psi_\theta(t) : \theta \in \Theta, t \in [0,\tau]\}$, where $\Theta$ is the class of all survival functions $t \mapsto \theta(t)$ with $\theta(0) = 0$ and with $t$ restricted to $[0,\tau]$, was shown to be Donsker. Note that in this setting, $\mathcal{H} = [0,\tau]$. Thus Conditions (B) and (C) of the theorem hold. It is also quite easy to verify directly that

$$\sup_{t\in[0,\tau]} P\left[\psi_\theta(t) - \psi_{\theta_0}(t)\right]^2 \to 0,$$

as $\|\theta - \theta_0\|_\infty \to 0$, and thus Condition (D) of the theorem is satisfied. If $\hat{\theta}_n$ is the Kaplan-Meier estimator, then $\|\Psi_n(\hat{\theta}_n)(t)\|_\infty = 0$ almost surely. If the bootstrapped version is

$$\hat{\theta}_n^\circ(t) \equiv \prod_{j:\tilde{T}_j \le t} \left(1 - \frac{n\mathbb{P}_n^\circ\left[\delta 1\{U = \tilde{T}_j\}\right]}{n\mathbb{P}_n^\circ 1\{U \ge \tilde{T}_j\}}\right),$$

where $\tilde{T}_1, \ldots, \tilde{T}_{m_n}$ are the observed failure times in the sample, then also $\|\Psi_n^\circ(\hat{\theta}_n^\circ)\|_\infty = 0$ almost surely. Thus Condition (E) of the theorem is satisfied, and hence all of the conditions of theorem are satisfied. Thus we obtain consistency, weak convergence, and bootstrap consistency for the Kaplan-Meier estimator all at once.

As mentioned at the end of Section 13.1, a master result such as this will not apply to all Z-estimator settings. Many interesting and important Z-estimators require an individualized approach to obtaining consistency, such as the Z-estimator for the proportional odds model for right-censored data which we examine in Chapter 15.

**Proof of Theorem 13.4.** The consistency of $\hat{\theta}_n$ and weak convergence of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ follow from Theorems 13.1 and 2.11 and Lemma 13.3. Theorem 13.1 also yields that there exists a decreasing sequence $0 < \eta_n \downarrow 0$ such that

$$\mathrm{P}\left(\|\hat{\theta}_n^\circ - \theta_0\| > \eta_n \middle| \mathcal{X}_n\right) = o_P(1).$$

Now we can use the same arguments used in the proof of Lemma 13.3, in combination with Theorem 2.6, to obtain that $\sqrt{n}(\Psi_n^\circ - \Psi)(\hat{\theta}_n^\circ) - \sqrt{n}(\Psi_n^\circ -$

$\Psi)(\theta_0) = E_n$, where $\mathrm{P}(E_n > \eta|\mathcal{X}_n) = o_P(1)$ for all $\eta > 0$. Combining this with arguments used in the proof of Theorem 2.11, we can deduce that $\sqrt{n}(\hat{\theta}_n^\circ - \theta_0) = -\dot{\Psi}_{\theta_0}^{-1}\sqrt{n}(\Psi_n^\circ - \Psi)(\theta_0) + E_n'$, where $\mathrm{P}(E_n' > \eta|\mathcal{X}_n) = o_P(1)$ for all $\eta > 0$. Combining this with the conclusion of Theorem 2.11, we obtain $\sqrt{n}(\hat{\theta}_n^\circ - \hat{\theta}_n) = -\dot{\Psi}_{\theta_0}^{-1}\sqrt{n}(\Psi_n^\circ - \Psi_n)(\theta_0) + E_n''$, where $\mathrm{P}(E_n'' > \eta|\mathcal{X}_n) = o_P(1)$ for all $\eta > 0$. The final conclusion now follows from reapplication of Theorem 2.6.□

## 13.3 Using the Delta Method

There is an alternative approach to Z-estimators that may be more effective for more general data settings, including non-i.i.d. and dependent data settings. The idea is to view the extraction of the zero from the estimating equation as a continuous mapping. Our approach is closely related to the approach in Section 3.9.4.7 of VW but with some important modifications which simplify the required assumptions. We require $\Theta$ to be the subset of a Banach space and $\mathbb{L}$ to be a Banach space. 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, and let $\Phi(\Theta, \mathbb{L})$ be the collection of all maps (or algorithms) $\phi : Z(\Theta, \mathbb{L}) \mapsto \Theta$ that for each element $z \in Z(\Theta, \mathbb{L})$ extract one of its zeros $\phi(z)$. This structure allows for multiple zeros.

The following lemma gives us a kind of uniform Hadamard differentiability of members of $\Phi(\Theta, \mathbb{L})$, which we will be able to use to obtain a delta method result for Z-estimators $\hat{\theta}_n$ that satisfy $\Psi_n(\hat{\theta}_n) = o_P(r_n^{-1})$ for some sequence $r_n \to \infty$ for which $X_n(\theta) \equiv r_n(\Psi_n - \Psi)(\theta)$ converges weakly to a tight, random element in $X \in \ell^\infty(\Theta_0, \mathbb{L})$, where $\Theta_0 \subset \Theta$ is an open neighborhood of $\theta_0$ and $\|X(\theta) - X(\theta_0)\|_{\mathbb{L}} \to 0$ as $\theta \to \theta_0$ almost surely, i.e., $X$ has continuous sample paths in $\theta$. Define $\ell_0^\infty(\Theta, \mathbb{L})$ to be the elements $x \in \ell^\infty(\Theta, \mathbb{L})$ for which $\|x(\theta) - x(\theta_0)\|_{\mathbb{L}} \to 0$ as $\theta \to \theta_0$.

THEOREM 13.5 *Assume $\Psi : \Theta \mapsto \mathbb{L}$ is uniformly norm-bounded over $\Theta$, $\Psi(\theta_0) = 0$, and Condition (13.1) holds. Let $\Psi$ also be Fréchet differentiable at $\theta_0$ with continuously invertible derivative $\dot{\Psi}_{\theta_0}$. Then the continuous linear operator $\phi'_\Psi : \ell_0^\infty(\Theta, \mathbb{L}) \mapsto lin\,\Theta$ defined by $z \mapsto \phi'_\Psi(z) \equiv -\dot{\Psi}_{\theta_0}^{-1}(z(\theta_0))$ satisfies:*

$$\sup_{\phi \in \Phi(\Theta, \mathbb{L})} \left\| \frac{\phi(\Psi + t_n z_n) - \phi(\Psi)}{t_n} - \phi'_\Psi(z(\theta_0)) \right\| \to 0,$$

*as $n \to \infty$, for any sequences $(t_n, z_n) \in (0, \infty) \times \ell^\infty(\Theta, \mathbb{L})$ such that $t_n \downarrow 0$, $\Psi + t_n z_n \in Z(\Theta, \mathbb{L})$, and $z_n \to z \in \ell_0^\infty(\Theta, \mathbb{L})$.*

**Proof.** Let $0 < t_n \downarrow 0$ and $z_n \to z \in \ell_0^\infty(\Theta, \mathbb{L})$ such that $\Psi + t_n z_n \in Z(\Theta, \mathbb{L})$. Choose any sequence $\{\phi_n\} \in \Phi(\Theta, \mathbb{L})$, and note that $\theta_n \equiv \phi_n(\Psi +$

$t_n Z_n$) satisfies $\Psi(\theta_n) + t_n z_n = 0$ by construction. Hence $\Psi(\theta_n) = O(t_n)$. By Condition (13.1), $\theta_n \to \theta_0$. By the Fréchet differentiability of $\Psi$,

$$\liminf_{n\to\infty} \frac{\|\Psi(\theta_n) - \Psi(\theta_0)\|_{\mathbb{L}}}{\|\theta_n - \theta_0\|} \geq \liminf_{n\to\infty} \frac{\|\dot{\Psi}_{\theta_0}(\theta_n - \theta_0)\|_{\mathbb{L}}}{\|\theta_n - \theta_0\|} \geq \inf_{\|g\|=1} \|\dot{\Psi}_{\theta_0}(g)\|_{\mathbb{L}},$$

where $g$ ranges over $\operatorname{lin}\Theta$. Since the inverse of $\dot{\Psi}_{\theta_0}$ is continuous, the right side of the above is positive. Thus there exists a universal constant $c < \infty$ (depending only on $\dot{\Psi}_{\theta_0}$ and $\operatorname{lin}\Theta$) for which $\|\theta_n - \theta_0\| < c\|\Psi(\theta_n) - \Psi(\theta_0)\|_{\mathbb{L}} = c\|t_n z_n(\theta_n)\|_{\mathbb{L}}$ for all $n$ large enough. Hence $\|\theta_n - \theta_0\| = O(t_n)$. By Fréchet differentiability, $\Psi(\theta_n) - \Psi(\theta_0) = \dot{\Psi}_{\theta_0}(\theta_n - \theta_0) + o(\|\theta_n - \theta_0\|)$, where $\dot{\Psi}_{\theta_0}$ is linear and continuous on $\operatorname{lin}\Theta$. The remainder term is $o(t_n)$ by previous arguments. Combining this with the fact that $t_n^{-1}(\Psi(\theta_n) - \Psi(\theta_0)) = -z_n(\theta_n) \to z(\theta_0)$, we obtain

$$\frac{\theta_n - \theta_0}{t_n} = \dot{\Psi}_{\theta_0}^{-1}\left(\frac{\Psi(\theta_n) - \Psi(\theta_0)}{t_n} + o(1)\right) \to -\dot{\Psi}_{\theta_0}^{-1}(z(\theta_0)).$$

The conclusion now follows since the sequence $\phi_n$ was arbitrary.□

The following simple corollary allows the delta method to be applied to Z-estimators. Let $\tilde{\phi} : \ell^\infty(\Theta, \mathbb{L}) \mapsto \Theta$ be a map such that for each $x \in \ell^\infty(\Theta, \mathbb{L})$, $\tilde{\phi}(x) = \theta_1 \neq \theta_0$ when $x \notin Z(\Theta, \mathbb{L})$ and $\tilde{\phi}(x) = \phi(x)$ for some $\phi \in \Phi(\Theta, \mathbb{L})$ otherwise.

Corollary 13.6 *Suppose $\Psi$ satisfies the conditions of Theorem 13.5, $\hat{\theta}_n = \tilde{\phi}(\Psi_n)$, and $\Psi_n$ has at least one zero for all $n$ large enough, outer almost surely. Suppose also that $r_n(\Psi_n - \Psi) \rightsquigarrow X$ in $\ell^\infty(\Theta, \mathbb{L})$, with $X$ tight and $\|X(\theta)\|_{\mathbb{L}} \to 0$ as $\theta \to \theta_0$ almost surely. Then $r_n(\hat{\theta}_n - \theta_0) \rightsquigarrow -\dot{\Psi}_{\theta_0}^{-1} X(\theta_0)$.*

**Proof.** Since $\Psi_n$ has a zero for all $n$ large enough, outer almost surely, we can, without loss of generality, assume that $\Psi_n$ has a zero for all $n$. Thus we can assume that $\tilde{\phi} \in \Phi(\Theta, \mathbb{L})$. By Theorem 13.5, we know that $\tilde{\phi}$ is Hadamard differentiable tangentially to $\ell_0^\infty(\Theta, \mathbb{L})$, which, by assumption, contains $X$ with probability 1. Thus Theorem 2.8 applies, and the desired result follows.□

We leave it as an exercise (see Exercise 13.4.2 below) to develop a corollary which utilizes $\tilde{\phi}$ to obtain a bootstrap result for Z-estimators. A drawback with this approach is that root finding algorithms in practice are seldom exact, and room needs to be allowed for computational error. The following corollary of Theorem 13.5 yields a very general Z-estimator result based on a modified delta method. We make the fairly realistic assumption that the Z-estimator $\hat{\theta}_n$ is computed from $\Psi_n$ using a deterministic algorithm (e.g., a computer program) that is allowed to depend on $n$ and which is not required to yield an exact root of $\Psi_n$.

Corollary 13.7 *Suppose $\Psi$ satisfies the conditions of Theorem 13.5, and $\hat{\theta}_n = A_n(\Psi_n)$ for some sequence of deterministic algorithms $A_n$ :*

*$\ell^\infty(\Theta,\mathbb{L}) \mapsto \Theta$ and random sequence $\Psi_n : \Theta \mapsto \mathbb{L}$ of estimating equations such that $\Psi_n \xrightarrow{\mathrm{P}} \Psi$ in $\ell^\infty(\Theta,\mathbb{L})$ and $\Psi_n(\hat\theta_n) = o_P(r_n^{-1})$, where $0 < r_n \to \infty$ is a sequence of constants for which $r_n(\Psi_n - \Psi) \rightsquigarrow X$ in $\ell^\infty(\Theta_0,\mathbb{L})$ for some closed $\Theta_0 \subset \Theta$ containing an open neighborhood of $\theta_0$, with $X$ tight and $\|X(\theta)\|_{\mathbb{L}} \to 0$ as $\theta \to \theta_0$ almost surely. Then $r_n(\hat\theta_n - \theta_0) \rightsquigarrow -\dot\Psi_{\theta_0}^{-1} X(\theta_0)$.*

**Proof.** Let $X_n \equiv r_n(\Psi_n - \Psi)$. By Theorem 7.26, there exists a new probability space $(\tilde\Omega, \tilde{\mathcal{A}}, \tilde P)$ on which: $\mathrm{E}^* f(\tilde\Psi_n) = \mathrm{E}^* f(\Psi_n)$ for all bounded $f : \ell^\infty(\Theta,\mathbb{L}) \mapsto \mathbb{R}$ and all $n \geq 1$; $\tilde\Psi_n \xrightarrow{\mathrm{as}^*} \Psi$ in $\ell^\infty(\Theta,\mathbb{L})$; $r_n(\tilde\Psi_n - \Psi) \xrightarrow{\mathrm{as}^*} \tilde X$ in $\ell^\infty(\Theta_0,\mathbb{L})$; $\tilde X$ and $X$ have the same distributions; and $r_n(A_n(\tilde\Psi_n) - \theta_0)$ and $r_n(A_n(\Psi_n) - \theta_0)$ have the same distributions. Note that for any bounded $f$, $\tilde\Psi_n \mapsto f(\tilde\Psi_n(A_n(\tilde\Psi_n))) = g(\tilde\Psi_n)$ for some bounded $g$. Thus $\tilde\Psi_n(\tilde\theta_n) = o_{\tilde P}(r_n^{-1})$ for $\tilde\theta_n \equiv A_n(\tilde\Psi_n)$.

Hence for any subsequence $n'$ there exists a further subsequence $n''$ such that $\tilde\Psi_{n''} \xrightarrow{\mathrm{as}^*} \Psi$ in $\ell^\infty(\Theta,\mathbb{L})$, $r_{n''}(\tilde\Psi_{n''} - \Psi) \xrightarrow{\mathrm{as}^*} \tilde X$ in $\ell^\infty(\Theta_0,\mathbb{L})$, and $\tilde\Psi_{n''}(\tilde\theta_{n''}) \xrightarrow{\mathrm{as}^*} 0$ in $\mathbb{L}$. Thus also $\Psi(\tilde\theta_{n''}) \xrightarrow{\mathrm{as}^*} 0$, which implies $\tilde\theta_{n''} \xrightarrow{\mathrm{as}^*} \theta_0$. Note that $\tilde\theta_{n''}$ is a zero of $\tilde\Psi_{n''}(\theta) - \tilde\Psi(\tilde\theta_{n''})$ by definition and is contained in $\Theta_0$ for all $n$ large enough. Hence, for all $n$ large enough, $r_{n''}(\tilde\theta_{n''} - \theta_0) = r_{n''}\left(\phi_{n''}\left(\tilde\Psi_{n''} - \tilde\Psi_{n''}(\tilde\theta_{n''})\right) - \phi_{n''}(\Psi)\right)$ for some sequence $\phi_n \in \Phi(\Theta_0,\mathbb{L})$ possibly dependent on sample realization $\tilde\omega \in \tilde\Omega$. This implies that for all $n$ large enough,

$$\begin{aligned}
&\left\| r_{n''}(\tilde\theta_{n''} - \theta_0) - \phi'_\Psi(\tilde X) \right\| \\
&\qquad \leq \sup_{\phi \in \Phi(\Theta_0,\mathbb{L})} \left\| r_{n''}\left(\phi\left(\tilde\Psi_{n''} - \tilde\Psi_{n''}(\tilde\theta_{n''})\right) - \phi(\Psi)\right) - \phi'_\Psi(\tilde X) \right\| \\
&\qquad \xrightarrow{\mathrm{as}^*} 0,
\end{aligned}$$

by Theorem 13.5 (with $\Theta_0$ replacing $\Theta$). This implies $\left\| r_{n''}(A_{n''}(\tilde\Psi_{n''}) - \theta_0) - \phi'_\Psi(\tilde X) \right\| \xrightarrow{\mathrm{as}^*} 0$. Since this holds for every subsequence, we have $r_n(A_n(\tilde\Psi_n) - \theta_0) \rightsquigarrow \phi'_\Psi(\tilde X)$. This of course implies $r_n(A_n(\Psi_n) - \theta_0) \rightsquigarrow \phi'_\Psi(X)$.□

The following corollary extends the previous result to generalized bootstrapped processes. Let $\Psi_n^\circ$ be a bootstrapped version of $\Psi_n$ based on both the data sequence $X_n$ (the data used in $\Psi_n$) and a sequence of weights $W = \{W_n, n \geq 1\}$.

Corollary 13.8 *Assume the conditions of Corollary 13.7 and, in addition, that $\hat\theta_n^\circ = A_n(\Psi_n^\circ)$ for a sequence of bootstrapped estimating equations $\Psi_n^\circ(X_n, W_n)$, with $\Psi_n^\circ - \Psi \underset{W}{\overset{\mathrm{P}}{\rightsquigarrow}} 0$ and $r_n\Psi_n^\circ(\hat\theta_n^\circ) \underset{W}{\overset{\mathrm{P}}{\rightsquigarrow}} 0$ in $\ell^\infty(\Theta,\mathbb{L})$, and with $r_n c(\Psi_n^\circ - \Psi_n) \underset{W}{\overset{\mathrm{P}}{\rightsquigarrow}} X$ in $\ell^\infty(\Theta_0,\mathbb{L})$ for some $0 < c < \infty$, where the maps $W_n \mapsto h(\Psi_n^\circ)$ are measurable for every $h \in C_b(\ell^\infty(\Theta,\mathbb{L}))$ outer almost surely. Then $r_n c(\hat\theta_n^\circ - \hat\theta_n) \underset{W}{\overset{\mathrm{P}}{\rightsquigarrow}} \phi'_\Psi(X)$.*

**Proof.** This proof shares many similarities with the proof of Theorem 12.1. To begin with, by using the same arguments used in the beginning of that proof, we can obtain that, unconditionally,

$$r_n \begin{pmatrix} \Psi_n^\circ - \Psi \\ \Psi_n - \Psi \end{pmatrix} \rightsquigarrow \begin{pmatrix} c^{-1}X_1' + X_2' \\ X_2' \end{pmatrix}$$

in $\ell^\infty(\Theta_0, \mathbb{L})$, where $X_1'$ and $X_2'$ are two independent copies of $X$. We can also obtain that $\Psi_n^\circ \xrightarrow{\mathrm{P}} \Psi$ in $\ell^\infty(\Theta, \mathbb{L})$ unconditionally. Combining this with a minor adaptation of the above Corollary 13.7 (see Exercise 13.4.3 below) we obtain unconditionally that

$$r_n \begin{pmatrix} \hat{\theta}_n^\circ - \theta_0 \\ \hat{\theta}_n - \theta_0 \\ \Psi_n^\circ(\theta_0) - \Psi(\theta_0) \\ \Psi_n(\theta_0) - \Psi(\theta_0) \end{pmatrix} \rightsquigarrow \begin{pmatrix} \phi_\Psi'(c^{-1}X_1' + X_2') \\ \phi_\Psi'(X_2') \\ c^{-1}X_1'(\theta_0) + X_2'(\theta_0) \\ X_2'(\theta_0) \end{pmatrix}.$$

This implies two things. First,

$$r_n c \begin{pmatrix} \hat{\theta}_n^\circ - \hat{\theta}_n \\ (\Psi_n^\circ - \Psi_n)(\theta_0) \end{pmatrix} \rightsquigarrow \begin{pmatrix} \phi_\Psi'(X) \\ X(\theta_0) \end{pmatrix}$$

unconditionally, since $\phi_\Psi'$ is linear on $\mathbb{L}$. Second, the usual continuous mapping theorem now yields unconditionally that

$$r_n c(\hat{\theta}_n^\circ - \hat{\theta}_n) + \dot{\Psi}_{\theta_0}^{-1}(r_n c(\Psi_n^\circ - \Psi_n)(\theta_0)) \xrightarrow{\mathrm{P}} 0, \tag{13.8}$$

since the map $(x, y) \mapsto x + \dot{\Psi}_{\theta_0}^{-1}(y)$ is continuous on all of $\operatorname{lin}\Theta \times \mathbb{L}$ (recall that $x \mapsto \phi_\Psi'(x) = -\dot{\Psi}_{\theta_0}^{-1}(x(\theta_0))$).

Now for any map $h \in C_b(\mathbb{L})$, the map $x \mapsto h(r_n c(x - \Psi_n(\theta_0)))$ is continuous and bounded for all $x \in \mathbb{L}$ outer almost surely. Thus the maps $W_n \mapsto h(r_n c(\Psi_n^\circ - \Psi_n)(\theta_0))$ are measurable for every $h \in C_b(\mathbb{L})$ outer almost surely. Hence the bootstrap continuous mapping theorem, Theorem 10.8, yields that $\dot{\Psi}_{\theta_0}^{-1}(r_n c(\hat{\Psi}_n^\circ - \Psi_n)(\theta_0)) \overset{\mathrm{P}}{\underset{W}{\rightsquigarrow}} \dot{\Psi}_{\theta_0}^{-1}(X)$. The desired result now follows from (13.8).□

An interesting application of the above results is to estimating equations for empirical processes from dependent data as discussed in Section 11.6. Specifically, suppose $\Psi_n(\theta)(h) = \mathbb{P}_n \psi_{\theta,h}$, where the stationary sample data $X_1, X_2, \ldots$ and $\mathcal{F} = \{\psi_{\theta,h} : \theta \in \Theta, h \in \mathcal{H}\}$ satisfy the conditions of Theorem 11.24 with marginal distribution $P$, and let $\Psi(\theta)(h) = P\psi_{\theta,h}$. Then the conclusion of Theorem 11.24 is that $\sqrt{n}(\Psi_n - \Psi) \rightsquigarrow \mathbb{H}$ in $\ell^\infty(\Theta \times \mathcal{H})$, where $\mathbb{H}$ is a tight, mean zero Gaussian process. Provided $\Psi$ satisfies the conditions of Theorem 13.1, and provided a few other conditions hold, Corollary 13.7 will give us weak convergence of the standardized Z-estimators $\sqrt{n}(\hat{\theta}_n - \theta_0)$ based on $\Psi_n$. Under certain regularity conditions, a moving blocks bootstrapped estimation equation $\Psi_n^\circ$ can be shown by Theorem 11.26 to satisfy

the requirements of Corollary 13.8. This enables valid bootstrap estimation of the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$. These results can also be extended to stationary sequences with long range dependence, where the normalizing rate $r_n$ may differ from $\sqrt{n}$.

## 13.4 Exercises

**13.4.1.** Show that the addition of $\bar{\xi}$ in the denominator of the weights in the weighted bootstrap introduced in Theorem 10.13, as discussed in Section 13.1, does not affect the conclusions of that theorem.

**13.4.2.** Develop a bootstrap central limit theorem for Z-estimators based Theorem 12.1 which utilizes the Hadamard differentiability of the zero-extraction map $\tilde{\phi}$ used in Corollary 13.6.

**13.4.3.** Verify the validity of the "minor adaptation" of Corollary 13.7 used in the proof of Corollary 13.8.

## 13.5 Notes

Theorem 2.11 and Lemma 13.2 are essentially a decomposition of Theorem 3.3.1 of VW into two parts. Lemma 13.3 is Lemma 3.3.5 of VW.

# 14
# M-Estimators

M-estimators, as introduced in Section 2.2.6, are approximate maximizers of objective functions computed from data. Note the estimators based on minimizing objective functions are trivially also M-estimators after taking the negative of the objective function. In some ways, M-estimators are more basic than Z-estimators since Z-estimators can always be expressed as M-estimators. The reverse is not true, however, since there are M-estimators that cannot be effectively formulated as Z-estimators. Nevertheless, Z-estimator theory is usually much easier to use whenever it can be applied. The focus, then, of this chapter is on M-estimator settings for which it is not practical to directly use Z-estimator theory. The usual issues for M-estimation theory are to establish consistency, determine the correct rate of convergence, establish weak convergence, and, finally, to conduct inference.

We first present a key result central to M-estimation theory, the argmax theorem, which permits deriving weak limits of M-estimators as the the argmax of the limiting process. This is useful for both consistency, which we discuss next, and weak convergence. The section on consistency includes a proof of Theorem 2.12 on Page 28. We then discuss how to determine the correct rate of convergence which is necessary for establishing weak convergence based on the argmax theorem. We then present general results for "regular estimators," i.e., estimators whose rate of convergence is $\sqrt{n}$. We then give several examples that illustrate M-estimation theory for non-regular estimators that have rates distinct from $\sqrt{n}$. Much of the content of this chapter is inspired by Chapter 3.2 of VW.

The M-estimators in both the regular and non-regular examples we present will be i.i.d. empirical processes of the form $M_n(\theta) = \mathbb{P}_n m_\theta$ for a class of measurable, real valued functions $\mathcal{M} = \{m_\theta : \theta \in \Theta\}$, where the parameter space $\Theta$ is usually a subset of a semimetric space. The entropy of the class $\mathcal{M}$ plays a crucial role in determining the proper rate of convergence. The aspect of the rate of convergence determining process is often the most difficult part technically in M-estimation theory and usually requires fairly precise bounds on moments of the empirical processes involved, such as those described in Section 11.1. We note that our presentation involves only a small amount of the useful results in the area. Much more of these kinds of results can be found in Chapter 3.4 of VW and in van de Geer (2000).

Inference for M-estimators is more challenging than it is for Z-estimators because the bootstrap is not in general automatically valid, especially when the convergence rate is not $\sqrt{n}$. On the other hand, subsampling $m$ out of $n$ observations (see Politis and Romano, 1994) can be shown to be universally valid, provided $m \to \infty$ and $m/n \to 0$. However, even this result is not entirely satisfactory because it requires $n$ to be quite large since $m$ must also be large yet small relative to $n$. Bootstrapping and other methods of inference for M-estimators is currently an area of active research, but we do not pursue it further in this chapter.

## 14.1 The Argmax Theorem

We now consider a sequence $\{M_n(h) : h \in H\}$ of stochastic processes indexed by a metric space $H$. Let $\hat{h}_n$ denote an M-estimator obtained by nearly-maximizing $M_n$. The idea of the argmax theorem presented below is that under reasonable regularity conditions, when $M_n \rightsquigarrow M$, where $M$ is another stochastic process in $\ell^\infty(H)$, that $\hat{h}_n \rightsquigarrow \hat{h}$, where $\hat{h}$ is the argmax of $M$. If we know that the rate of convergence of an M-estimator $\hat{\theta}_n$ is $r_n$ (a nondecreasing, positive sequence), where $\hat{\theta}_n$ is the argmax of $\theta \mapsto M_n(\theta)$, then $\hat{h}_n = r_n(\hat{\theta}_n - \theta_0)$ can be expressed as the argmax of $h \mapsto \tilde{M}_n(h) \equiv r_n [M_n(\theta_0 + h/r_n) - M_n(\theta_0)]$. Provided $\tilde{M}_n \rightsquigarrow M$, and the regularity conditions of the argmax theorem apply, $\hat{h}_n = r_n(\hat{\theta}_n - \theta_0) \rightsquigarrow \hat{h}$, where $\hat{h}$ is the argmax of $M$. Consistency results will follow for the choice $r_n = 1$ for all $n \geq 1$.

Note that in the theorem, we require the sequence $\hat{h}_n$ to be uniformly tight. This is stronger than asymptotic tightness, as pointed out in Lemma 7.10, but is also quite easy to establish for Euclidean parameters which will be our main focus in this chapter. For finite Euclidean estimators that are measurable, uniform tightness follows automatically from asymptotic tightness (see Exercise 14.6.1). This is a reasonable restriction, since, in practice, most infinite-dimensional estimators that converge weakly can usually be

expressed as Z-estimators. Our consistency results that we present later on will not require uniform tightness and will thus be more readily applicable to infinite dimensional estimators. Returning to the theorem at hand, most weak convergence results for non-regular estimators apply to finite dimensional parameters, and thus the theorem below will be applicable. We also note that it is not hard to modify these results for applicability to specific settings, including some infinite dimensional settings. For interesting examples in this direction, see Ma and Kosorok (2005a) and Kosorok and Song (2007). We now present the argmax theorem, which utilizes upper semicontinuity as defined in Section 6.1:

THEOREM 14.1 *(Argmax theorem) Let $M_n, M$ be stochastic processes indexed by a metric space $H$ such that $M_n \rightsquigarrow M$ in $\ell^\infty(K)$ for every compact $K \subset H$. Suppose also that almost all sample paths $h \mapsto M(h)$ 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 $M_n(\hat{h}_n) \geq \sup_{h \in H} M_n(h) - o_P(1)$, then $\hat{h}_n \rightsquigarrow \hat{h}$ in $H$.*

**Proof.** Fix $\epsilon > 0$. By uniform tightness of $\hat{h}_n$ and tightness of $\hat{h}$, there exists a compact set $K \subset H$ such that $\liminf_{n\to\infty} \mathrm{P}^*(\hat{h}_n \in K) \geq 1 - \epsilon$ and $\mathrm{P}(\hat{h} \in K) \geq 1 - \epsilon$. Then almost surely

$$M(\hat{h}) \quad > \quad \sup_{h \notin G, h \in K} M(h),$$

for every open $G \ni \hat{h}$, by upper semicontinuity of $M$. To see this, suppose it were not true. Then by the compactness of $K$, there would exist a convergent sequence $h_m \in G^c \cap K$, for some open $G \ni \hat{h}$, with $h_m \to h$ and $M(h_m) \to M(\hat{h})$. The upper semicontinuity forces $M(h) \geq M(\hat{h})$ which contradicts the uniqueness of the maximum.

Now apply Lemma 14.2 below for the sets $A = B = K$, to obtain

$$\begin{aligned}
\limsup_{n\to\infty} \mathrm{P}^*(\hat{h}_n \in F) &\leq \limsup_{n\to\infty} \mathrm{P}^*(\hat{h}_n \in F \cap K) + \limsup_{n\to\infty} \mathrm{P}^*(\hat{h}_n \notin K) \\
&\leq \mathrm{P}(\hat{h} \in F \cup K^c) + \epsilon \\
&\leq \mathrm{P}(\hat{h} \in F) + \mathrm{P}(\hat{h} \in K^c) + \epsilon \\
&\leq \mathrm{P}(\hat{h} \in F) + 2\epsilon.
\end{aligned}$$

The desired result now follows from the Portmanteau theorem since $\epsilon$ was arbitrary.□

LEMMA 14.2 *Let $M_n, M$ be stochastic processes indexed by a metric space $H$, and let $A, B \subset H$ be arbitrary. Suppose there exists a random element $\hat{h}$ such that almost surely*

$$M(\hat{h}) \quad > \quad \sup_{h \notin G, h \in K} M(h), \quad \textit{for every open } G \ni \hat{h}. \tag{14.1}$$

*Suppose the sequence $\hat{h}_n$ satisfies $M_n(\hat{h}_n) \geq \sup_{h\in H} M_n(h) - o_P(1)$. Then, if $M_n \rightsquigarrow M$ in $\ell^\infty(A \cup B)$, we have for every closed set $F$,*

$$\limsup_{n\to\infty} \mathrm{P}^*(\hat{h}_n \in F \cap A) \leq \mathrm{P}(\hat{h} \in F \cup B^c).$$

**Proof.** By the continuous mapping theorem,

$$\sup_{h\in F\cap A} M_n(h) - \sup_{h\in B} M_n(h) \quad \rightsquigarrow \quad \sup_{h\in F\cap A} M(h) - \sup_{h\in B} M(h),$$

and thus

$$\begin{aligned}
&\limsup_{n\to\infty} \mathrm{P}^*(\hat{h}_n \in F \cap A) \\
&\qquad \leq \limsup_{n\to\infty} \mathrm{P}^* \left( \sup_{h\in F\cap A} M_n(h) \geq \sup_{h\in H} M_n(h) - o_P(1) \right) \\
&\qquad \leq \limsup_{n\to\infty} \mathrm{P}^* \left( \sup_{h\in F\cap A} M_n(h) \geq \sup_{h\in B} M_n(h) - o_P(1) \right) \\
&\qquad \leq \mathrm{P} \left( \sup_{h\in F\cap A} M(h) \geq \sup_{h\in B} M(h) \right),
\end{aligned}$$

by Slutsky's theorem (to get rid of the $o_P(1)$ part) followed by the Portmanteau theorem. Note that the event $E$ in the last probability can't happen when $\hat{h} \in F^c \cap B$ because of Assumption (14.1) and the fact that $F^c$ is open. Thus $E$ is contained in the set $\{\hat{h} \in F\} \cup \{\hat{h} \notin B\}$, and the conclusion of the lemma follows.□

## 14.2 Consistency

We can obtain a consistency result by specializing the argmax theorem to the setting where $M$ is fixed. This will not yield as general a result as Theorem 2.12 because of the uniform tightness requirement. The primary goal of this section is to prove Theorem 2.12 on Page 28. Before giving the proof, we want to present a result comparing a few different ways of establishing identifiability. We assume throughout this section that $(\Theta, d)$ is a metric space. In the following lemma, the condition given in (i) is the identifiability condition assumed in Theorem 2.12, while the Condition (ii) is often called the "well-separated maximum" condition:

LEMMA 14.3 *Let $M : \Theta \mapsto \mathbb{R}$ be a map and $\theta_0 \in \Theta$ a point. The following conditions are equivalent:*

*(i) For any sequence $\{\theta_n\} \in \Theta$, $\liminf_{n\to\infty} M(\theta_n) \geq M(\theta_0)$ implies $d(\theta_n, \theta_0) \to 0$.*

*(ii) For every open* $G \ni \theta_0$, $M(\theta_0) > \sup_{\theta \notin G} M(\theta)$.

*The following condition implies both (i) and (ii):*

*(iii)* $M$ *is upper semicontinuous with a unique maximum at* $\theta_0$.

**Proof.** Suppose (i) is true but (ii) is not. Then there exists an open $G \ni \theta_0$ such that $\sup_{\theta \notin G} \geq M(\theta_0)$. This implies the existence of a sequence $\theta_n$ with both $\liminf_{n\to\infty} M(\theta_n) \geq M(\theta_0)$ and $d(\theta_n, \theta_0) \to \tau > 0$, which is a contradiction. Thus (i) implies (ii). Now assume (ii) is true but (i) is not. Then there exists a sequence with $\liminf_{n\to\infty} M(\theta_n) \geq M(\theta_0)$ but with $\theta_n \notin G$ for all $n$ large enough and some open $G \ni \theta_0$. Of course, this contradicts (ii), and thus (ii) implies (i). Now suppose $M$ is upper semicontinuous with a unique maximum at $\theta_0$ but (ii) does not hold. Then there exists an open $G \ni \theta_0$ for which $\sup_{\theta \notin G} M(\theta) \geq M(\theta_0)$. But this implies that the set $\{\theta : M(\theta) \geq M(\theta_0)\}$ contains at least one point in addition to $\theta_0$ since $G^c$ is closed. This contradiction completes the proof.□

Any one of the three identifiability conditions given in the above lemma are sufficient for Theorem 2.12. The most convenient condition in practice will depend on the setting. Here is the awaited proof:

**Proof of Theorem 2.12 (Page 28).** Since $\liminf_{n\to\infty} M(\theta_n) \geq M(\theta_0)$ implies $d(\theta_n, \theta_0) \to 0$ for any sequence $\{\theta_n\} \in \Theta$, we know that there exists a non-decreasing cadlag function $f : [0, \infty] \mapsto [0, \infty]$ that satisfies both $f(0) = 0$ and $d(\theta, \theta_0) \leq f(|M(\theta) - M(\theta_0)|)$ for all $\theta \in \Theta$. The details for constructing such an $f$ are left as an exercise (see Exercise 14.6.2).

For Part (i), note that $M(\theta_0) \geq M(\hat{\theta}_n) \geq M_n(\hat{\theta}_n) - \|M_n - M\|_\Theta \geq M_n(\theta_0) - o_P(1) \geq M(\theta_0) - o_P(1)$. By the previous paragraph, this implies $d(\hat{\theta}_n, \theta_0) \leq f(|M(\hat{\theta}_n) - M(\theta_0)|) \xrightarrow{\mathrm{P}} 0$. An almost identical argument yields Part (ii).□

## 14.3 Rate of Convergence

In this section, we relax the requirement that $(\Theta, d)$ be a metric space to only requiring that it to be a semimetric space. If $\theta \mapsto M(\theta)$ is two times differentiable at a point of maximum $\theta_0$, then the first derivative of $M$ at $\theta_0$ must vanish while the second derivative should be negative definite. Thus it is not unreasonable to require that $M(\theta) - M(\theta_0) \leq -cd^2(\theta, \theta_0)$ for all $\theta$ in a neighborhood of $\theta_0$ and some $c > 0$. The following theorem shows that an upper bound for the rate of convergence of a near-maximizer of a random objection function $M_n$ can be obtained from the modulus of continuity of $M_n - M$ at $\theta_0$. In practice, one may need to try several rates that satisfy the conditions of this theorem before finding the right $r_n$ for which the weak limit of $r_n(\hat{\theta}_n - \theta_0)$ is nontrivial.

THEOREM 14.4 *(Rate of convergence) Let $M_n$ be a sequence of stochastic processes indexed by a semimetric space $(\Theta, d)$ and $M : \Theta \mapsto \mathbb{R}$ a deterministic function such that for every $\theta$ in a neighborhood of $\theta_0$, there exists a $c_1 > 0$ such that*

$$M(\theta) - M(\theta_0) \quad \leq \quad -c_1 \tilde{d}^2(\theta, \theta_0), \tag{14.2}$$

*where $\tilde{d} : \Theta \times \Theta \mapsto [0, \infty)$ satisfies $\tilde{d}(\theta_n, \theta_0) \to 0$ whenever $d(\theta_n, \theta_0) \to 0$. Suppose that for all $n$ large enough and sufficiently small $\delta$, the centered process $M_n - M$ satisfies*

$$\mathrm{E}^* \sup_{\tilde{d}(\theta,\theta_0)<\delta} \sqrt{n}\,|(M_n - M)(\theta) - (M_n - M)(\theta_0)| \quad \leq \quad c_2 \phi_n(\delta), \tag{14.3}$$

*for $c_2 < \infty$ and functions $\phi_n$ such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ not depending on $n$. Let*

$$r_n^2 \phi_n(r_n^{-1}) \leq c_3 \sqrt{n}, \quad \text{for every } n \text{ and some } c_3 < \infty. \tag{14.4}$$

*If the sequence $\hat{\theta}_n$ satisfies $M_n(\hat{\theta}_n) \geq \sup_{\theta\in\Theta} M_n(\theta) - O_P(r_n^{-2})$ and converges in outer probability to $\theta_0$, then $r_n \tilde{d}(\hat{\theta}_n, \theta_0) = O_P(1)$.*

**Proof.** We will use a modified "peeling device" (see, for example, Section 5.3 of van de Geer, 2000) for the proof. For every $\eta > 0$, let $\eta' > 0$ be a number for which $\tilde{d}(\theta, \theta_0) \leq \eta$ whenever $\theta \in \Theta$ satisfies $d(\theta, \theta_0) \leq \eta'$ and also $\tilde{d}(\theta, \theta_0) \leq \eta/2$ whenever $\theta \in \Theta$ satisfies $d(\theta, \theta_0) \leq \eta'/2$. Such an $\eta'$ always exists for each $\eta$ by the assumed relationship between $d$ and $\tilde{d}$. Note also that $M_n(\hat{\theta}_n) \geq \sup_{\theta\in\Theta} M_n(\theta) - O_P(r_n^{-2}) \geq M_n(\theta_0) - O_P(r_n^{-2})$. Now fix $\epsilon > 0$, and choose $K < \infty$ such that the probability that $M_n(\hat{\theta}_n) - M_n(\theta_0) < -Kr_n^{-2}$ is $\leq \epsilon$.

For each $n$, the parameter space minus the point $\theta_0$ can be partitioned into "peels" $S_{j,n} = \{\theta : 2^{j-1} < r_n \tilde{d}(\theta, \theta_0) \leq 2^j\}$ with $j$ ranging over the integers. Assume that $M_n(\hat{\theta}_n) - M_n(\theta_0) \geq -Kr_n^{-2}$, and note that if $r_n \tilde{d}(\hat{\theta}_n, \theta_0)$ is $> 2^M$ for a given integer $M$, then $\hat{\theta}_n$ is in one of the peels $S_{j,n}$, with $j > M$. In that situation, the supremum of the map $\theta \mapsto M_n(\theta) - M_n(\theta_0) + Kr_n^{-2}$ is nonnegative by the property of $\hat{\theta}_n$. Conclude that for every $\eta > 0$,

$$\begin{aligned} &\mathrm{P}^* \left( r_n \tilde{d}(\hat{\theta}_n, \theta_0) > 2^M \right) \\ &\quad \leq \sum_{j \geq M, 2^j \leq \eta r_n} \mathrm{P}^* \left( \sup_{\theta \in S_{j,n}} \left[ M_n(\theta) - M_n(\theta_0) + Kr_n^{-2} \right] \geq 0 \right) \\ &\qquad + \mathrm{P}^* \left( 2d(\hat{\theta}_n, \theta_0) \geq \eta' \right) + \mathrm{P}^* \left( M_n(\hat{\theta}_n) - M_n(\theta_0) < -Kr_n^{-2} \right). \end{aligned} \tag{14.5}$$

The $\limsup_{n\to\infty}$ of the sum of the two probabilities after the summation on the right side is $\leq \epsilon$ by the consistency of $\hat{\theta}_n$ and the choice of $K$. Now

choose $\eta$ small enough so that (14.2) holds for all $d(\theta, \theta_0) \leq \eta'$ and (14.3) holds for all $\delta \leq \eta$. Then for every $j$ involved in the sum, we have for every $\theta \in S_{j,n}$, $M(\theta) - M(\theta_0) \leq -c_1 \tilde{d}^2(\theta, \theta_0) \leq -c_1 2^{2j-2} r_n^{-2}$. In terms of the centered process $W_n \equiv M_n - M$, the summation on the right side of (14.5) may be bounded by

$$\begin{aligned} \sum_{j \geq M, 2^j \leq \eta r_n} \mathrm{P}^* \left( \|W_n(\theta) - W_n(\theta_0)\|_{S_{j,n}} \geq \frac{c_1 2^{2j-2} - K}{r_n^2} \right) & \\ \leq \sum_{j \geq M} \frac{c_2 \phi_n(2^j / r_n) r_n^2}{\sqrt{n}(c_1 2^{2j-2} - K)} & \\ \leq \sum_{j \geq M} \frac{c_2 c_3 2^{j\alpha}}{(c_1 2^{2j-2} - K)}, & \end{aligned}$$

by Markov's inequality, the conditions on $r_n$, and the fact that $\phi_n(c\delta) \leq c^\alpha \phi_n(\delta)$ for every $c > 1$ as a consequence of the assumptions on $\phi_n$. It is not difficult to show that this last sum goes to zero as $M \to \infty$ (verifying this is saved as an exercise). Thus we can choose an $M < \infty$ such that the $\limsup_{n \to \infty}$ of the left side of (14.5) is $\leq 2\epsilon$. The desired result now follows since $\epsilon$ was arbitrary.□

Consider the i.i.d. setting with criterion functions of the form $M_n(\theta) = \mathbb{P}_n m_\theta$ and $M(\theta) = P m_\theta$. The scaled and centered process $\sqrt{n}(M_n - M) = \mathbb{G}_n m_\theta$ equals the empirical process at $m_\theta$. The Assertion (14.3) involves assessing the suprema of the empirical process by classes of functions $\mathcal{M}_\delta \equiv \{m_\theta - m_{\theta_0} : \tilde{d}(\theta, \theta_0) < \delta\}$. Taking this view, establishing (14.3) will require fairly precise—but not unreasonably precise—knowledge of the involved empirical process. The moment results in Section 11.1 will be useful here, and we will illustrate this with several examples later on in this chapter. We note that the problems we address in this book represent only a small subset of the scope and capabilities of empirical process techniques for determining rates of M-estimators. We close this section with the following corollary which essentially specializes Theorem 14.4 to the i.i.d. setting. Because the specialization is straightforward, we omit the somewhat trivial proof. Recall that the relation $a \lesssim b$ means that $a$ is less than or equal $b$ times a universal finite and positive constant.

COROLLARY 14.5 *In the i.i.d. setting, assume that for every $\theta$ in a neighborhood of $\theta_0$, $P(m_\theta - m_{\theta_0}) \lesssim -\tilde{d}^2(\theta, \theta_0)$, where $\tilde{d}$ satisfies the conditions given in Theorem 14.4. Assume moreover that there exists a function $\phi$ such that $\delta \mapsto \phi(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ and, for every $n$, $\mathrm{E}^* \|\mathbb{G}_n\|_{\mathcal{M}_\delta} \lesssim \phi(\delta)$. If the sequence $\hat{\theta}_n$ satisfies $\mathbb{P}_n m_{\hat{\theta}_n} \geq \sup_{\theta \in \Theta} \mathbb{P}_n m_\theta - O_P(r_n^{-2})$ and converges in outer probability to $\theta_0$, then $r_n \tilde{d}(\hat{\theta}_n, \theta_0) = O_P(1)$ for every sequence $r_n$ for which $r_n^2 \phi(1/r_n) \lesssim \sqrt{n}$ for all $n \geq 1$.*

## 14.4 Regular Euclidean M-Estimators

A general result for Euclidean M-estimators based on i.i.d. data was given in Theorem 2.13 on Page 29 of Section 2.2.6. We now prove this theorem. In Section 2.2.6, the theorem was used to establish asymptotic normality of a least-absolute-deviation regression estimator. Establishing asymptotic normality in this situation is quite difficult without empirical process methods.

**Proof of Theorem 2.13 (Page 29).** We first utilize Corollary 14.5 to verify that $\sqrt{n}$ is the correct rate of convergence. We will use Euclidean distance as both the discrepancy measure and distance through, i.e. $\tilde{d}(\theta_1, \theta_2) = d(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|$. Condition (2.18) of the theorem indicates that $\mathcal{M}_\delta$ in this instance is a Lipschitz class, and thus Theorem 9.23 implies that

$$N_{[]}(2\epsilon\|F_\delta\|_{P,2}, \mathcal{M}_\delta, L_2(P)) \lesssim \epsilon^{-p}, \tag{14.6}$$

where $F_\delta \equiv \delta\dot{m}$ is an envelope for $\mathcal{M}_\delta$. To see this, it may be helpful to rewrite Condition (2.18) as

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq \frac{\|\theta_1 - \theta_2\|}{\delta} F_\delta(x).$$

Now (14.6) can be utilized in Theorem 11.2 to obtain

$$\mathrm{E}^*\|\mathbb{G}_n\|_{\mathcal{M}_\delta} \lesssim \|F_\delta\|_{P,2} \lesssim \delta.$$

Hence the modulus of continuity condition in Corollary 14.5 is satisfied for $\phi(\delta) = \delta$. Combining Condition (2.19), the maximality of $\theta_0$, and the fact that the second derivative matrix $V$ is nonsingular and continuous, yields that $M(\theta) - M(\theta_0) \lesssim -\|\theta - \theta_0\|^2$. Since $\phi(\delta)/\delta^\alpha = \delta^{1-\alpha}$ is decreasing for any $\alpha \in (1, 2)$ and $n\phi(1/\sqrt{n}) = n/\sqrt{n} = \sqrt{n}$, the remaining conditions of Corollary 14.5 are satisfied for $r_n = \sqrt{n}$, and thus $\sqrt{n}(\hat{\theta}_n - \theta_0) = O_P(1)$.

The next step is to apply the argmax theorem (Theorem 14.1) to the process $h \mapsto U_n(h) \equiv n(M_n(\theta_0 + h/\sqrt{n}) - M_n(\theta_0))$. Now fix a compact $K \subset \mathbb{R}^p$, and note that

$$\begin{aligned} U_n(h) &= \mathbb{G}_n\left[\sqrt{n}\left(m_{\theta_0 + h/\sqrt{n}} - m_{\theta_0}\right) - h^T\dot{m}_{\theta_0}\right] \\ &\quad + h^T\mathbb{G}_n\dot{m}_{\theta_0} + n(M(\theta_0 + h/\sqrt{n}) - M(\theta_0)) \\ &\equiv E_n(h) + h^T\mathbb{G}_n\dot{m}_{\theta_0} + \frac{1}{2}h^TVh + o(1), \end{aligned}$$

where $o(1)$ denotes a quantity going to zero uniformly over $K$. Note that $\hat{h}_n \equiv \sqrt{n}(\hat{\theta}_n - \theta_0)$ satisfies $U_n(\hat{h}_n) \geq \sup_{h\in\mathbb{R}^p} U_n(h) - o_P(1)$. Thus, provided we can establish that $\|E_n\|_K = o_P(1)$, the argmax theorem will yield that $\hat{h}_n \rightsquigarrow \hat{h}$, where $\hat{h}$ is the argmax of $h \mapsto U(h) \equiv h^TZ + (1/2)h^TVh$, where

$Z$ is the Gaussian limiting distribution of $\mathbb{G}_n \dot{m}_{\theta_0}$. Hence $\hat{h} = -V^{-1}Z$ and the desired result will follow.

We now prove $\|E_n\|_K = o_P(1)$ for all compact $K \subset \mathbb{R}^p$. let $u_h^n(x) \equiv \sqrt{n}(m_{\theta_0+h/\sqrt{n}}(x) - m_{\theta_0}(x)) - h^T \dot{m}_{\theta_0}(x)$, and note that by (2.18),

$$|u_{h_1}^n(x) - u_{h_2}^n(x)| \le (\dot{m}(x) + \|\dot{m}_{\theta_0}(x)\|)\|h_1 - h_2\|,$$

for all $h_1, h_2 \in \mathbb{R}^p$ and all $n \ge 1$. Fix a compact $K \subset \mathbb{R}^p$, and let $\mathcal{F}_n \equiv \{u_h^n : h \in K\}$. Applying Theorem 9.23 once again, but with $\|\cdot\| = \|\cdot\|_{Q,2}$ (for any probability measure $Q$ on $\mathcal{X}$) instead of $\|\cdot\|_{P,2}$, we obtain

$$N_{[]}(2\epsilon\|F_n\|_{Q,2}, \mathcal{F}_n, L_2(Q)) \le k\epsilon^{-p},$$

where the envelope $F_n \equiv (\dot{m} + \|\dot{m}_{\theta_0}\|)\|h\|_K$ and $k < \infty$ does not depend on $K$ or $n$. Lemma 9.18 in Chapter 9 now yields that

$$\sup_{n\ge 1} \sup_Q N(\epsilon\|F_n\|_{Q,2}, \mathcal{F}_n, L_2(Q)) \le k\left(\frac{2}{\epsilon}\right)^p,$$

where the second supremum is taken over all finitely discrete probability measures on $\mathcal{X}$. This implies that Condition (A) of Theorem 11.20 holds for $\mathcal{F}_n$ and $F_n$. In addition, Condition (2.19) implies that Condition (B) of Theorem 11.20 also holds with $H(s,t) = 0$ for all $s, t \in K$. It is not difficult to verify that all of the remaining conditions of Theorem 11.20 also hold (we save this as an exercise), and thus $\mathbb{G}_n u_h^n \rightsquigarrow 0$ in $\ell^\infty(K)$. This, of course, is the desired result.□

## 14.5 Non-Regular Examples

We now present two examples in detail that illustrate the techniques presented in this chapter for parameter estimation with non-regular rates of convergence. The first example considers a simple change-point model with three parameters wherein two of the parameter estimates converge at the regular rate while one of the parameter estimates converges at the $n$-rate, i.e., it converges faster than $\sqrt{n}$. The second example is monotone density estimation based on the Grenander estimator which is shown to yield convergence at the cube-root rate.

### *14.5.1 A Change-Point Model*

For this model, we observe i.i.d. realizations of $X = (Y, Z)$, where $Y = \alpha 1\{Z \le \zeta\} + \beta 1\{Z > \zeta\} + \epsilon$, $Z$ and $\epsilon$ are independent with $\epsilon$ continuous, $\mathrm{E}\epsilon = 0$ and $\sigma^2 \equiv \mathrm{E}\epsilon^2 < \infty$, $\gamma \equiv (\alpha, \beta) \in \mathbb{R}^2$ and $\zeta$ is known to lie in a bounded interval $[a, b]$. The unknown parameters can be collected as $\theta = (\gamma, \zeta)$, and the subscript zero will be used to denote the true parameter

values. We make the very important assumption that $\alpha_0 \neq \beta_0$ and also assume that $Z$ has a strictly bounded and positive density $f$ over $[a, b]$ with $\mathrm{P}(Z < a) > 0$ and $\mathrm{P}(Z > b) > 0$. Our goal is to estimate $\theta$ through least squares. This is the same as maximizing $M_n(\theta) = \mathbb{P}_n m_\theta$, where

$$m_\theta(x) \equiv -\left(y - \alpha 1\{z \leq \zeta\} - \beta 1\{z > \zeta\}\right)^2.$$

Let $\hat{\theta}_n$ be maximizers of $M_n(\theta)$, where $\hat{\theta}_n \equiv (\hat{\gamma}_n, \hat{\zeta}_n)$ and $\hat{\gamma}_n \equiv (\hat{\alpha}_n, \hat{\beta}_n)$.

Since we are not assuming that $\gamma$ is bounded, we first need to prove the existence of $\hat{\gamma}_n$, i.e., we need to prove that $\|\hat{\gamma}_n\| = O_P(1)$. We then need to provide consistency of all parameters and then establish the rates of convergence for the parameters. Finally, we need to obtain the joint limiting distribution of the parameter estimates.

*Existence.*

Note that the covariate $Z$ and parameter $\zeta$ can be partitioned into four mutually exclusive sets: $\{Z \leq \zeta \wedge \zeta_0\}$, $\{\zeta < Z \leq \zeta_0\}$, $\{\zeta_0 < Z \leq \zeta\}$ and $\{Z > \zeta \vee \zeta_0\}$. Since also $1\{Z < a\} \leq 1\{Z \leq \zeta \wedge \zeta_0\}$ and $1\{Z > b\} \leq 1\{Z > \zeta \vee \zeta_0\}$ by assumption, we obtain $-\mathbb{P}_n \epsilon^2 = M_n(\theta_0) \leq M_n(\hat{\theta}_n)$

$$\leq -\mathbb{P}_n \left[(\epsilon - \hat{\alpha}_n + \alpha_0)^2 1\{Z < a\} + (\epsilon - \hat{\beta}_n + \beta_0)^2 1\{Z > b\}\right].$$

By decomposing the squares, we now have

$$\begin{aligned}
(\hat{\alpha}_n - \alpha_0)^2 \mathbb{P}_n[1\{Z < a\}] + (\hat{\beta}_n - \beta_0)^2 \mathbb{P}_n[1\{Z > b\}] & \\
\leq \quad & \mathbb{P}_n[\epsilon^2 1\{a \leq z \leq b\}] \\
& +2|\hat{\alpha}_n - \alpha_0| \mathbb{P}_n[\epsilon 1\{Z < a\}] + 2|\hat{\beta}_n - \beta_0| \mathbb{P}_n[\epsilon 1\{Z > b\}] \\
\leq \quad & O_p(1) + o_P(1)\|\hat{\gamma}_n - \gamma_0\|.
\end{aligned}$$

Since $\mathrm{P}(Z < a) \wedge \mathrm{P}(Z > b) > 0$, the above now implies that $\|\hat{\gamma}_n - \gamma_0\|^2 = O_P(1 + \|\hat{\gamma}_n - \gamma_0\|)$ and hence that $\|\hat{\gamma}_n - \gamma_0\| = O_P(1)$. Thus all the parameters are bounded in probability and therefore exist.

*Consistency.*

Our approach to establishing consistency will be to utilize the argmax theorem (Theorem 14.1). We first need to establish that $M_n \rightsquigarrow M$ in $\ell^\infty(K)$ for all compact $K \subset H \equiv \mathbb{R}^2 \times [a, b]$, where $M(\theta) \equiv P m_\theta$. We then need to show that $\theta \mapsto M(\theta)$ is upper semicontinuous with a unique maximum at $\theta_0$. We already know from the previous paragraph that $\hat{\theta}_n$ is asymptotically tight (i.e., $\|\hat{\theta}_n\| = O_P(1)$). The argmax theorem will then yield that $\hat{\theta}_n \rightsquigarrow \theta_0$ as desired.

Fix a compact $K \subset H$. We now verify that $\mathcal{F}_K \equiv \{m_\theta : \theta \in K\}$ is Glivenko-Cantelli. Note that

$$\begin{aligned} m_\theta(X) &= -(\epsilon-\alpha+\alpha_0)^2 1\{Z \le \zeta \wedge \zeta_0\} - (\epsilon-\beta+\alpha_0)^2 1\{\zeta < Z \le \zeta_0\} \\ &\quad -(\epsilon-\alpha+\beta_0)^2 1\{\zeta_0 < Z \le \zeta\} - (\epsilon-\beta+\beta_0)^2 1\{Z > \zeta \vee \zeta_0\}. \end{aligned}$$

It is not difficult to verify that $\{(\epsilon-\alpha+\alpha_0)^2 : \theta \in K\}$ and $1\{Z \le \zeta \wedge \zeta_0 : \theta \in K\}$ are separately Glivenko-Cantelli classes. Thus the product of the two class is also Glivenko-Cantelli by Corollary 9.27 since the product of the two envelopes is integrable. Similar arguments reveal that the remaining components of the sum are also Glivenko-Cantelli, and reapplication of Corollary 9.27 yields that $\mathcal{F}_K$ itself is Glivenko-Cantelli. Thus $M_n \rightsquigarrow M$ in $\ell^\infty(K)$ for all compact $K$.

We now establish upper semicontinuity of $\theta \mapsto M(\theta)$ and uniqueness of the maximum. Using the decomposition of the sets for $(Z, \zeta)$ used in the *Existence* paragraph above, we have

$$\begin{aligned} M(\theta) &= -\mathrm{P}\epsilon^2 - (\alpha-\alpha_0)^2 \mathrm{P}(Z \le \zeta \wedge \zeta_0) - (\beta-\alpha_0)^2 \mathrm{P}(\zeta < Z \le \zeta_0) \\ &\quad -(\alpha-\beta_0)^2 \mathrm{P}(\zeta_0 < Z \le \zeta) - (\beta-\beta_0)^2 \mathrm{P}(Z > \zeta \vee \zeta_0) \\ &\le -\mathrm{P}\epsilon^2 = M(\theta_0). \end{aligned}$$

Because $Z$ has a bounded density on $[a, b]$, we obtain that $M$ is continuous. It is also clear that $M$ has a unique maximum at $\theta_0$ because the density of $Z$ is bounded below and $\alpha_0 \neq \beta_0$ (see Exercise 14.6.5 below). Now the conditions of the argmax theorem are met, and the desired consistency follows.

*Rate of convergence.*

We will utilize Corollary 14.5 to obtain the convergence rates via the discrepancy function $\tilde{d}(\theta, \theta_0) \equiv \|\gamma - \gamma_0\| + \sqrt{|\zeta - \zeta_0|}$. Note that this is not a norm since it does not satisfy the triangle inequality. Nevertheless, $\tilde{d}(\theta, \theta_0) \to 0$ if and only if $\|\theta - \theta_0\| \to 0$. Moreover, from the *Consistency* paragraph above, we have that

$$\begin{aligned} M(\theta) - M(\theta_0) &= -P\{Z \le \zeta \wedge \zeta_0\}(\alpha-\alpha_0)^2 - P\{Z > \zeta \vee \zeta_0\}(\beta-\beta_0)^2 \\ &\quad -P\{\zeta < Z \le \zeta_0\}\}(\beta-\alpha_0)^2 \\ &\quad -P\{\zeta_0 < Z \le \zeta\}(\alpha-\beta_0)^2 \\ &\le -P\{Z < a\}(\alpha-\alpha_0)^2 - P\{Z > b\}(\beta-\beta_0)^2 \\ &\quad -k_1(1-o(1))|\zeta-\zeta_0| \\ &\le -(k_1 \wedge \delta_1 - o(1))\tilde{d}^2(\theta, \theta_0), \end{aligned}$$

where the first inequality follows from the fact that the product of the density of $Z$ and $(\alpha_0 - \beta_0)^2$ is bounded below by some $k_1 > 0$, and the second inequality follows from both $\mathrm{P}(Z < a)$ and $\mathrm{P}(Z > b)$ being bounded below by some $\delta_1 > 0$. Thus $M(\theta) - M(\theta_0) \lesssim -\tilde{d}^2(\theta, \theta_0)$ for all $\|\theta - \theta_0\|$ small enough, as desired.

Consider now the class of functions $\mathcal{M}_\delta \equiv \{m_\theta - m_{\theta_0} : \tilde{d}(\theta, \theta_0) < \delta\}$. Using previous calculations, we have

(14.7)

$$\begin{aligned} m_\theta - m_{\theta_0} &= 2(\alpha-\alpha_0)\epsilon 1\{Z \le \zeta \wedge \zeta_0\} + 2(\beta-\beta_0)\epsilon 1\{Z > \zeta \vee \zeta_0\} \\ &\quad +2(\beta-\alpha_0)\epsilon 1\{\zeta < Z \le \zeta_0\} + 2(\alpha-\beta_0)\epsilon 1\{\zeta_0 < Z \le \zeta\} \\ &\quad -(\alpha-\alpha_0)^2 1\{Z \le \zeta \wedge \zeta_0\} - (\beta-\beta_0)^2 1\{Z > \zeta \vee \zeta_0\} \\ &\quad -(\beta-\alpha_0)^2 1\{\zeta < Z \le \zeta_0\} - (\alpha-\beta_0)^2 1\{\zeta_0 < Z \le \zeta\} \\ &\equiv A_1(\theta) + A_2(\theta) + B_1(\theta) + B_2(\theta) \\ &\quad -C_1(\theta) - C_2(\theta) - D_1(\theta) - D_2(\theta). \end{aligned}$$

Consider first $A_1$. Since $\{1\{Z \le t\} : t \in [a,b]\}$ is a VC class, it is easy to compute that

$$\mathrm{E}^* \sup_{\tilde{d}(\theta,\theta_0)<\delta} |\mathbb{G}_n A_1(\theta)| \lesssim \delta,$$

as a consequence of Lemma 8.17. Similar calculations apply to $A_2$. Similar calculations also apply to $C_1$ and $C_2$, except that the upper bounds will be $\lesssim \delta^2$ instead of $\lesssim \delta$. Now we consider $B_1$. An envelope for the class $\mathcal{F} = \{B_1(\theta) : \tilde{d}(\theta,\theta_0) < \delta\}$ is $F = 2(|\beta_0 - \alpha_0| + \delta)|\epsilon| 1\{\zeta_0 - \delta^2 < Z \le \zeta_0\}$. It is not hard to verify that

$$\log N_{[]}(\eta \|F\|_{P,2}, \mathcal{F}, L_2(P)) \lesssim \log(1/\eta) \tag{14.8}$$

(see Exercise 14.6.6). Now Theorem 11.2 yields that

$$\mathrm{E}^* \sup_{\tilde{d}(\theta,\theta_0)<\delta} |\mathbb{G}_n B_1(\theta)| = \mathrm{E}^* \|\mathbb{G}_n\|_{\mathcal{F}} \times \|F\|_{P,2} \lesssim \delta^2.$$

Similar calculations apply also to $B_2$, $D_1$ and $D_2$. Combining all of these results with the fact that $O(\delta^2) = O(\delta)$, we obtain $\mathrm{E}^* \|\mathbb{G}_n\|_{\mathcal{M}_\delta} \lesssim \delta$.

Now when $\delta \mapsto \phi(\delta) = \delta$, $\phi(\delta)/\delta^\alpha$ is decreasing for any $\alpha \in (1,2)$. Thus the conditions of Corollary 14.5 are satisfied with $\phi(\delta) = \delta$. Since $r_n^2 \phi(1/r_n) = r_n$, we obtain that $\sqrt{n}\tilde{d}(\hat{\theta}_n, \theta_0) = O_P(1)$. By the form of $\tilde{d}$, this now implies that $\sqrt{n}\|\hat{\gamma}_n - \gamma_0\| = O_P(1)$ and $n|\hat{\zeta}_n - \zeta_0| = O_P(1)$.

*Weak convergence.*

We will utilize a minor modification of the argmax theorem and the rate result above to obtain the limiting distribution of $\hat{h}_n = (\sqrt{n}(\hat{\gamma}_n - \gamma_0), n(\hat{\zeta}_n - \zeta_0))$. From the rate result, we know that $\hat{h}_n$ is uniformly tight and is the smallest argmax of $h \mapsto Q_n(h) \equiv n\mathbb{P}_n(m_{\theta_{n,h}} - m_{\theta_0})$, where $\theta_{n,h} \equiv \theta_0 + (h_1/\sqrt{n}, h_2/\sqrt{n}, h_3/n)$ and $h \equiv (h_1,h_2,h_3) \in \mathbb{R}^3 \equiv H$. Note that we have qualified $\hat{h}_n$ as being the smallest argmax, which is interpreted componentwise since $H$ is three dimensional. This is because if we hold $(h_1,h_2)$ fixed, $M_n(\theta_{n,h})$ does not vary in $h_3$ over the interval $n[Z_{(j)} - \zeta_0, Z_{(j+1)} - \zeta_0)$,

for $j = 0, \ldots, n$, where $Z_{(1)}, \ldots, Z_{(n)}$ are the order statistics for $Z_1, \ldots, Z_n$, $Z_{(0)} \equiv -\infty$, and $Z_{(n+1)} \equiv \infty$. Because $\mathrm{P}(Z < a) > 0$, we only need to consider $h_3$ at the values $n(Z_{(j)} - \zeta_0)$, $j = 1, \ldots, n$, provided $n$ is large enough.

Let $\mathcal{D}_K$ be the space of functions $q : K \subset H \mapsto \mathbb{R}$, that are continuous in the first two arguments $(h_1, h_2)$ and right-continuous and piecewise constant in the third argument $h_3$, with jump locations constant across $(h_1, h_2)$, where $K$ is a closed and bounded, rectangular subset of $H$ that contains an open neighborhood of zero. For each $q \in \mathcal{D}_K$, let $h_3 \mapsto J_q(h_3)$ be the cadlag counting process denoting the jump locations with $J_q(0-) = 0$, jumps of size positive 1 at each jump point in $q(\cdot, \cdot, h_3)$ for $h_3 \geq 0$, and with $h_3 \mapsto J_q(-h_3)$ also having jumps of size positive 1 (but left-continuous) at each jump point in $q(\cdot, \cdot, h_3)$ also for $h_3 \geq 0$ (the left-continuity comes from the reversed time scale). Thus $J_q(h_3)$ is decreasing for $h_3 < 0$ and increasing for $h_3 \geq 0$.

For $q_1, q_2 \in \mathcal{D}_K$, define the distance $d_K(q_1, q_2)$ to be the sum of a "modified Skorohod metric" $\tilde{d}_K(q_1, q_2)$ (to be defined shortly) and the Skorohod distance between $J_{q_1}$ and $J_{q_2}$. Note that the rectangular structure of $K$ ensures that $K = K_1 \times K_2 \times K_3$, where each $K_j$ is a closed interval in $\mathbb{R}$, $j = 1, 2, 3$. For each closed interval $C$ in $\mathbb{R}$, define $\Lambda_C$ to be the collection of continuous, strictly increasing maps $\lambda : C \mapsto C$ such that $\lambda(C) = C$. Define a norm on $\Lambda_C$ to be

$$\lambda \mapsto \|\lambda\| \quad \equiv \quad \sup_{s \neq t : s,t \in C} \log \left| \frac{\lambda(t) - \lambda(s)}{t - s} \right|. \tag{14.9}$$

For $q_1, q_2 \in \mathcal{D}_K$, we define

$$\tilde{d}_K(q_1, q_2) \equiv \inf_{\lambda \in \Lambda_{K_3}} \left\{ \sup_{h \in K} |q_1(h_1, h_2, h_3) - q_2(h_1, h_2, \lambda(h_3))| + \|\lambda\| \right\}.$$

We save the verification that $\tilde{d}_K$ is a metric as an exercise (see Exercise 14.6.7).

Now it is not difficult to see that the smallest argmax function is continuous on $\mathcal{D}_K$ with respect to $d_K$. We will argue that $Q_n(h) \rightsquigarrow Q(h)$ in $(\mathcal{D}_K, d_K)$, for some limiting process $Q$, and for each compact $K \subset H$. By the continuous mapping theorem, the smallest argmax of the restriction of $Q_n$ to $K$ will converge weakly to the smallest argmax of the restriction of $Q$ to $K$. Since $\hat{h}_n$ is uniformly tight, we obtain $\hat{h}_n \rightsquigarrow \hat{h}$, where $\hat{h}$ is the smallest argmax of $Q$.

All that remains is to establish the specified weak convergence and to characterize $Q$. We first argue that $Q_n - \tilde{Q}_n = o_P^K(1)$ in $(\mathcal{D}_K, d_K)$ for each compact $K \subset H$, where

$$\begin{aligned}\tilde{Q}_n(h) &\equiv 2h_1\sqrt{n}\mathbb{P}_n[\epsilon 1\{Z\le\zeta_0\}] - h_1^2\mathrm{P}(Z\le\zeta_0)\\ &\quad +2h_2\sqrt{n}\mathbb{P}_n[\epsilon 1\{Z>\zeta_0\}] - h_2^2\mathrm{P}(Z>\zeta_0)\\ &\quad +n\mathbb{P}_n[-2(\alpha_0-\beta_0)\epsilon-(\alpha_0-\beta_0)^2]1\{\zeta_0+h_3/n<Z\le\zeta_0\}\\ &\quad +n\mathbb{P}_n[2(\alpha_0-\beta_0)\epsilon-(\alpha_0-\beta_0)^2]1\{\zeta_0<Z\le\zeta_0+h_3/n\}\\ &\equiv \tilde{A}_n(h_1)+\tilde{B}_n(h_2)+\tilde{C}_n(h_3)+\tilde{D}_n(h_3).\end{aligned}$$

The superscript $K$ in $o_P^K(1)$ indicates that the error is in terms of $d_k$. Fix a compact $K\subset H$. Note that by (14.7), $\tilde{A}_n(h_1)=n\mathbb{P}_n[A_1(\theta_{n,h})-C_1(\theta_{n,h}]+\tilde{E}_n(h)$, where

$$\begin{aligned}\tilde{E}_n(h) &= 2h_1\mathbb{G}_n\left[1\{Z\le(\zeta_0+h_3/n)\wedge\zeta_0\}-1\{Z\le\zeta_0\}\right]\\ &\quad -h_1^2\left[\mathbb{P}_n 1\{Z\le(\zeta_0+h_3/n)\wedge\zeta_0\}-\mathrm{P}(Z\le\zeta_0)\right]\\ &\to 0\end{aligned}$$

in probability, as $n\to\infty$, uniformly over $h\in K$. A similar analysis reveals the uniform equivalence of $\tilde{B}_n(h_2)$ and $n\mathbb{P}_n[A_2(\theta_{n,h}-C_2(\theta_{n,h})]$. It is fairly easy to see that $\tilde{C}_n(h_3)$ and $n\mathbb{P}_n[B_1(\theta_{n,h})-D_1(\theta_{n,h})]$ are asymptotically uniformly equivalent in probability as also $\tilde{D}_n(h_3)$ and $n\mathbb{P}_n[B_2(\theta_{n,h})-D_2(\theta_{n,h})]$. Thus $Q_n-\tilde{Q}_n$ goes to zero, in probability, uniformly over $h\in K$. Note that the potential jump points in $h_3$ for $Q_n$ and $\tilde{Q}_n$ remain the same, and thus $Q_n-\tilde{Q}_n=o_P^K(1)$ as desired.

Lemma 14.6 below shows that $\tilde{Q}_n \rightsquigarrow Q\equiv 2h_1Z_1-h_1^2\mathrm{P}(Z\le\zeta_0)+2h_2Z_2-h_2^2\mathrm{P}(Z>\zeta_0)+Q^+(h_3)1\{h_3>0\}+Q^-(h_3)1\{h_3<0\}$ in $(\mathcal{D}_K,d_K)$, where $Z_1$, $Z_2$, $Q^+$ and $Q^-$ are all independent and $Z_1$ and $Z_2$ are mean zero Gaussian with respective variances $\sigma^2\mathrm{P}(Z\le\zeta_0)$ and $\sigma^2\mathrm{P}(Z>\zeta_0)$. Let $s\mapsto\nu^+(s)$ be a right-continuous homogeneous Poisson process on $[0,\infty)$ with intensity parameter $f(\zeta_0)$ (recall that $f$ is the density of $\epsilon$), and let $s\mapsto\nu^-(s)$ be another Poisson process, independent of $\nu^+$, on $[-\infty,0)$ which is left-continuous and goes backward in time with intensity $f(\zeta_0)$. Let $(V_k^+)_{k\ge1}$ and $(V_k^-)_{k\ge1}$ be independent sequences of i.i.d. random variables with $V_1^+$ being a realization of $2(\alpha_0-\beta_0)\epsilon-(\alpha_0-\beta_0)^2$ and $V_1^-$ being a realization of $-2(\alpha_0-\beta_0)\epsilon-(\alpha_0-\beta_0)^2$. Also define $V_0^+=V_0^-=0$ for convenience. Then $h_3\mapsto Q^+(h_3)\equiv 1\{h_3>0\}\sum_{0\le k\le\nu^+(h_3)}V_k^+$ and $h_3\mapsto Q^-(h_3)\equiv 1\{h_3<0\}\sum_{0\le k\le\nu^-(h_3)}V_k^-$.

Putting this all together, we conclude that $\hat{h}=(\hat{h}_1,\hat{h}_2,\hat{h}_3)$, where all three components are mutually independent, $\hat{h}_1$ and $\hat{h}_2$ are both mean zero Gaussian with respective variances $\sigma^2/\mathrm{P}(Z\le\zeta_0)$ and $\sigma^2/\mathrm{P}(Z>\zeta_0)$, and where $\hat{h}_3$ is the smallest argmax of $h_3\mapsto Q^+(h_3)+Q^-(h_3)$. Note that the expected value of both $V_1^+$ and $V_1^-$ is $-(\alpha_0-\beta_0)^2$. Thus $Q^++Q^-$ will be zero at $h_3=0$ and eventually always negative for all $h_3$ far enough away from zero. This means that the smallest argmax of $Q^++Q^-$ will be bounded in probability as desired.

LEMMA 14.6 *For each closed, bounded, rectangular $K \subset H$ that contains an open neighborhood of zero, $\tilde{Q}_n \rightsquigarrow Q$ in $(\mathcal{D}_K, d_K)$.*

**Proof.** Fix $K$. The approach we take will be to establish asymptotic tightness of $\tilde{Q}_n$ on $\mathcal{D}_K$ and convergence of all finite dimensional distributions. More precisely, we need $(\tilde{Q}_n, J_{\tilde{Q}_n}) \rightsquigarrow (Q, \nu^- + \nu^+)$ in the product of the "modified Skorohod" and Skorohod topologies. This boils down to establishing

$$\begin{pmatrix} \sqrt{n}\mathbb{P}_n\epsilon 1\{Z \le \zeta_0\} \\ \sqrt{n}\mathbb{P}_n\epsilon 1\{Z > \zeta_0\} \\ n\mathbb{P}_n 1\{\zeta_0 + h_3/n < Z \le \zeta_0\} \\ n\mathbb{P}_n\epsilon 1\{\zeta_0 + h_3/n < Z \le \zeta_0\} \\ n\mathbb{P}_n 1\{\zeta_0 < Z \le \zeta_0 + h_3/n\} \\ n\mathbb{P}_n\epsilon 1\{\zeta_0 < Z \le \zeta_0 + h_3/n\} \end{pmatrix} \rightsquigarrow \begin{pmatrix} Z_1 \\ Z_2 \\ \nu^- \\ Q^- \\ \nu^+ \\ Q^+ \end{pmatrix}, \tag{14.10}$$

in the product of the $\mathbb{R}^2$ and $(D[-K, K])^{\otimes 4}$ topologies, where $D[-K, K]$ is the space of cadlag functions on $[-K, K]$ endowed with the Skorohod topology.

It can be shown by an adaptation of the characteristic function approach used in the proof of Theorem 5 of Kosorok and Song (2007), that all finite dimensional distributions in (14.10) converge. We will omit the somewhat lengthy arguments.

We next need to show that the processes involved are asymptotically tight. Accordingly, consider first $\tilde{R}_n(u) \equiv n\mathbb{P}_n 1\{\zeta_0 < Z \le \zeta_0 + u/n\}$, and note that $E\tilde{R}_n(K) = Kf(\zeta_0) < \infty$. Thus the number of jumps in the cadlag, monotone increasing, piece-wise constant process $\tilde{R}_n$ over $[0, K]$ is bounded in probability. Fix $\eta \in (0, K]$, and let $0 = u_0 < u_1 < \cdots u_m = K$ be a finite partition of $[0, K]$ such that $m < 2/\eta$ with $\max_{1 \le j \le m} |u_j - u_{j-1}| < \eta$. Note that the limiting probability that any two jumps in $\tilde{R}_n$ occur within the interval $(u_{j-1}, u_j]$ is bounded above by $\eta^2 f^2(\zeta_0)$. Thus, for arbitrary $\eta > 0$, the limiting probability that any two jumps in $\tilde{R}_n$ occur within a distance $\eta$ anywhere in $[0, K]$ is $O(\eta)$. Hence

$$\lim_{\eta \downarrow 0} \limsup_{n \to \infty} \mathrm{P}\left( \sup_{s,t \in [0,K]: |s-t| < \eta} |\tilde{R}_n(s) - \tilde{R}_n(t)| > \eta \right) = 0,$$

and $\tilde{R}_n$ is therefore asymptotically tight. Similar arguments can be used to show that the remaining processes are also asymptotically tight. The desired conclusion now follows.□

### *14.5.2 Monotone Density Estimation*

This example is a special case of the cube root asymptotic results of Kim and Pollard (1990) which was analyzed in detail in Section 3.2.14 of VW.

Let $X_1, \ldots, X_n$ be a sample of size $n$ 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 non-increasing step function equal to the left derivative of the *least concave majorant* of the empirical distribution function $\mathbb{F}_n$. This $\hat{f}_n$ is the celebrated Grenander estimator (Grenander, 1956). For a fixed value of $t > 0$, we will study the properties of $\hat{f}_n(t)$ under the assumption that $f$ is differentiable at $t$ with derivative $-\infty < f'(t) < 0$. Specifically, we will establish consistency of $\hat{f}_n(t)$, verify that the rate of convergence of $\hat{f}_n$ is $n^{1/3}$, and derive weak convergence of $n^{1/3}(\hat{f}_n(t) - f(t))$. Existence of $\hat{f}_n$ will be verified automatically as a consequence of consistency.

*Consistency.*

Let $\hat{F}_n$ denote the least concave majorant of $\mathbb{F}_n$. In general, the least concave majorant of a function $g$ is the smallest concave function $h$ such that $h \geq g$. One can construct $\hat{F}_n$ by imagining a string tied at $(x, y) = (0, 0)$ which is pulled tight over the top of the function graph $(x, y = \mathbb{F}_n(x))$. The slope of each of the piecewise linear segments will be non-increasing, and the string $(\hat{F}_n)$ will touch $\mathbb{F}_n$ at two or points $(x_j, y_j)$, $j = 0, \ldots, k$, where $k \geq 1$, $(x_0, y_0) \equiv (0, 0)$ and $x_k$ is the last observation in the sample. For all $x > x_k$, we set $\hat{F}_n(x) = 1$. Note also that $\hat{F}_n$ is continuous. We leave it as an exercise to verify that this algorithm does indeed produce the least concave majorant of $\mathbb{F}_n$. The following lemma (Marshall's lemma) yields that $\hat{F}_n$ is uniformly consistent for $F$. We save the proof as another exercise.

LEMMA 14.7 *(Marshall's lemma) Under the give conditions,* $\sup_{t \geq 0} |\hat{F}_n(t) - F(t)| \leq \sup_{t \geq 0} |\mathbb{F}_n(t) - F(t)|$.

Now fix $0 < \delta < t$, and note that by definition of $\hat{f}_n$,

$$\frac{\hat{F}_n(t + \delta) - \hat{F}_n(t)}{\delta} \leq \hat{f}_n(t) \leq \frac{\hat{F}_n(t) - \hat{F}_n(t - \delta)}{\delta}.$$

By Marshall's lemma, the upper and lower bounds converge almost surely to $\delta^{-1}(F(t) - F(t - \delta))$ and $\delta^{-1}(F(t + \delta) - F(t))$, respectively. By the assumptions on $F$ and the arbitrariness of $\delta$, we obtain $\hat{f}_n(t) \stackrel{\text{as*}}{\to} f(t)$.

*Rate of convergence.*

To determine the rate of convergence, we need to perform an interesting inverse transformation of the problem that will also be useful for obtaining the weak limiting distribution. Define the stochastic process $\{\hat{s}_n(a) : a > 0\}$ by $\hat{s}_n(a) = \text{argmax}_{s \geq 0}\{\mathbb{F}_n(s) - as\}$, where the largest value is selected when multiple maximizers exist. The function $\hat{s}_n$ is a sort of 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 $t \geq 0$ and $a > 0$. To see this, first assume that $\hat{f}_n(t) \leq a$. This means that the left derivative of $\hat{F}_n$ is $\leq a$ at $t$. Hence a line of slope $a$

that is moved down vertically from $+\infty$ will first touch $\hat{F}_n$ at a point $s_0$ to the left of (or equal to) $t$. That point is also the point at which $\hat{F}_n$ is furthest away from the line $s \mapsto as$ passing through the origin. Thus $s_0 = \text{argmax}_{s \geq 0}\{\mathbb{F}_n(s) - as\}$, and hence $\hat{s}_n(a) \leq t$. Now suppose $\hat{s}_n(a) \leq t$. Then the argument can be taken in reverse to see that the slope of the line that touches $\hat{F}_n$ at $\hat{s}_n(a)$ is less than or equal to the left derivative of $\hat{F}_n$ at $t$, and thus $\hat{f}_n(t) \leq a$. Hence,

$$\text{(14.11)} \quad \mathrm{P}(n^{1/3}(\hat{f}_n(t) - f(t)) \leq x) \quad = \quad \mathrm{P}(\hat{s}_n(f(t) + xn^{-1/3}) \leq t),$$

and the desired rate and weak convergence result can be deduced from the argmax values of $x \mapsto \hat{s}_n(f(t) + xn^{-1/3})$. Applying the change of variable $s \mapsto t + g$ in the definition of $\hat{s}_n$, we obtain

$$\hat{s}_n(f(t) + xn^{-1/3}) - t = \text{argmax}_{\{g > -t\}}\{\mathbb{F}_n(t+g) - (f(t) + xn^{-1/3})(t+g)\}.$$

In this manner, the probability on the left side of (14.11) is precisely $\mathrm{P}(\hat{g}_n \leq 0)$, where $\hat{g}_n$ is the argmax above.

Now, by the previous argmax expression combined with the fact that the location of the maximum of a function does not change when the function is shifted vertically, we have $\hat{g}_n \equiv \text{argmax}_{\{g > -t\}}\{M_n(g) \equiv \mathbb{F}_n(t+g) - \mathbb{F}_n(t) - f(t)g - xgn^{-1/3}\}$. It is not hard to see that $\hat{g}_n = O_P(1)$ and that $M_n(g) \xrightarrow{\mathrm{P}} M(g) \equiv F(t+g) - F(t) - f(t)g$ uniformly on compacts, and thus $\hat{g}_n = o_P(1)$. We now utilize Theorem 14.4 to obtain the rate for $\hat{g}_n$, with the metric $d(\theta_1, \theta_2) = |\theta_1 - \theta_2|$, $\theta = g$, $\theta_0 = 0$ and $\tilde{d} = d$. Note the fact that $M_n(0) = M(0) = 0$ will simplify the calculations. It is now easy to see that $M(g) \lesssim -g^2$, and by using Theorem 11.2, that

$$\begin{aligned} \mathrm{E}^* \sup_{|g| < \delta} \sqrt{n}\,|M_n(g) - M(g)| \quad &\leq \quad \mathrm{E}^* \sup_{|g| < \delta} |\mathbb{G}_n(1\{X \leq t+g\} - 1\{X \leq t\})| \\ &\quad + O(\sqrt{n}\delta n^{-1/3}) \\ &\lesssim \quad \phi_n(\delta) \equiv \delta^{1/2} + \sqrt{n}\delta n^{-1/3}. \end{aligned}$$

Clearly, $\phi_n(\delta)/\delta^\alpha$ is decreasing in $\delta$ for $\alpha = 3/2$. Since $n^{2/3}\phi_n(n^{-1/3}) = n^{1/2} + n^{1/6}n^{-1/3} = O(n^{1/2})$, Theorem 14.4 yields $n^{1/3}\hat{g}_n = O_P(1)$. We show in the next section how this enables weak convergence of $n^{1/3}(\hat{f}(t) - f(t))$.

*Weak convergence.*

Let $\hat{h}_n = n^{1/3}\hat{g}_n$, and note that since the maximum of a function does not change when the function is multiplied by a constant, we have that $\hat{h}_n$ is the argmax of the process

$$\begin{aligned} \text{(14.12)} \quad h \mapsto\ & n^{2/3} M_n(n^{-1/3}h) \\ =\ & n^{2/3}(\mathbb{P}_n - P)\left(1\{X \leq t + hn^{-1/3}\} - 1\{X \leq t\}\right) \\ & + n^{2/3}\left[F(t + hn^{-1/3}) - F(t) - f(t)hn^{-1/3}\right] - xh. \end{aligned}$$

Fix $0 < K < \infty$, and apply Theorem 11.20 to the sequence of classes $\mathcal{F}_n = \{n^{1/6}\left(1\{X \le t + hn^{-1/3}\} - 1\{X \le t\}\right) : -K \le h \le K\}$ with envelope sequence $F_n = n^{1/6}1\{t - Kn^{-1/3} \le X \le t + Kn^{-1/3}\}$, to obtain that the process on the right side of (14.12) converges in $\ell^\infty(-K, K)$ to

$$h \mapsto \mathbb{H}(h) \equiv \sqrt{f(t)}\mathbb{Z}(h) + \frac{1}{2}f'(t)h^2 - xh,$$

where $\mathbb{Z}$ is a two-sided Brownian motion originating at zero (two independent Brownian motions starting at zero, one going to the right of zero and the other going to the left). From the previous paragraph, we know that $\hat{h}_n = O_P(1)$. Since it is not hard to verify that $\mathbb{H}$ is continuous with a unique maximum, the argmax theorem now yields by the arbitrariness of $K$ that $\hat{h}_n \rightsquigarrow \hat{h}$, where $\hat{h} = \operatorname{argmax}\mathbb{H}$. By Exercise 14.6.10 below, we can simplify the form of $\hat{h}$ to

$$\left|\frac{4f(t)}{[f'(t)]^2}\right|^{1/3} \operatorname{argmax}_h\{\mathbb{Z}(h) - h^2\} + \frac{x}{f'(t)}.$$

Since

$$\begin{aligned} &\mathrm{P}\left(\left|\frac{4f(t)}{[f'(t)]^2}\right|^{1/3} \operatorname{argmax}_h\{\mathbb{Z}(h) - h^2\} + \frac{x}{f'(t)} \le 0\right) \\ &\qquad = \mathrm{P}\left(|4f'(t)f(t)|^{1/3} \operatorname{argmax}_h\left\{\mathbb{Z} - h^2\right\} \le x\right), \end{aligned}$$

we now have, as a result of (14.11), that

$$n^{1/3}(\hat{f}_n(t) - f(t)) \rightsquigarrow |f'(t)f(t)|^{1/3}\mathbb{C},$$

where the random variable $\mathbb{C} \equiv \operatorname{argmax}_h\{\mathbb{Z}(h) - h^2\}$ has Chernoff's distribution (see Groeneboom, 1989).

## 14.6 Exercises

**14.6.1.** Show that for a sequence $X_n$ of measurable, Euclidean random variables which are finite almost surely, measurability plus asymptotic tightness implies uniform tightness.

**14.6.2.** For a metric space $(\mathbb{D}, d)$, let $H : \mathbb{D} \mapsto [0, \infty]$ be a function such that $H(x_0) = 0$ for a point $x_0 \in \mathbb{D}$ and $H(x_n) \to 0$ implies $d(x_n, x_0) \to 0$ for any sequence $\{x_n\} \in \mathbb{D}$. Show that there exists a non-decreasing cadlag function $f : [0, \infty] \mapsto [0, \infty]$ that satisfies both $f(0) = 0$ and $d(x, x_0) \le f(|H(x)|)$ for all $x \in H$. Hint: Use the fact that the given conditions on $H$ imply the existence of a decreasing sequence $0 < \tau_n \downarrow 0$ such that $H(x) < \tau_n$ implies $d(x, x_0) < 1/n$, and note that it is permissible to have $f(u) = \infty$ for all $u \ge \tau_1$.

**14.6.3.** In the proof of Theorem 14.4, verify that for fixed $c < \infty$ and $\alpha < 2$,

$$\sum_{j \geq M} \frac{2^{j\alpha}}{2^{2j} - c} \to 0,$$

as $M \to \infty$.

**14.6.4.** In the context of the last paragraph of the proof of Theorem 2.13, given in Section 14.4, complete the verification of the conditions of Theorem 11.20.

**14.6.5.** Consider the function $\theta \mapsto M(\theta)$ defined in Section 14.5.1. Show that it has a unique maximum over $\mathbb{R}^2 \times [a, b]$. Also show that the maximum is not unique if $\alpha_0 = \beta_0$.

**14.6.6.** Verify (14.8).

**14.6.7.** Show that the modified Skorohod metric $\tilde{d}_K$ defined in Section 14.5.1 is a metric. Hint: First show that for any $\lambda_1, \lambda_2 \in \Lambda_C$, we have

$$\|\lambda_1(\lambda_2)\| \leq \|\lambda_1\| + \|\lambda_2\|,$$

for the norm defined in (14.9).

**14.6.8.** Verify that the algorithm described in the second paragraph of Section 14.5.2 does indeed generate the least concave majorant of $\mathbb{F}_n$.

**14.6.9.** The goal of this exercise is to prove Marshall's lemma given in Section 14.5.2. Denoting $A_n(t) \equiv \hat{F}_n(t) - F(t)$ and $B_n(t) \equiv \mathbb{F}_n(t) - F(t)$, the proof can be broken into the following steps:

(a) Show that $0 \geq \inf_{t \geq 0} A_n(t) \geq \inf_{t \geq 0} B_n(t)$.

(b) Show that

i. $\sup_{t \geq 0} A_n(t) \geq 0$ and $\sup_{t \geq 0} B_n(t) \geq 0$.
ii. If $\sup_{t \geq 0} B_n(t) = 0$, then $\sup_{t \geq 0} A_n(t) = 0$.
iii. If $\sup_{t \geq 0} B_n(t) > 0$, then $\sup_{t \geq 0} A_n(t) \leq \sup_{t \geq 0} B_n(t)$ (this last step is tricky).

Now verify that $0 \leq \sup_{t \geq 0} A_n(t) \leq \sup_{t \geq 0} B_n(t)$.

(c) Now complete the proof.

**14.6.10.** Let $\{\mathbb{Z}(h) : h \in \mathbb{R}\}$ be a standard two-sided Brownian motion with $\mathbb{Z}(0) = 0$. (The process is zero-mean Gaussian and the increment $\mathbb{Z}(g) - \mathbb{Z}(h)$ has variance $|g - h|$.) Then $\operatorname{argmax}_h\{a\mathbb{Z}(h) - bh^2 - ch\}$ is equal in distribution to $(a/b)^{2/3}\operatorname{argmax}_g\{\mathbb{Z}(g) - g^2\} - c/(2b)$, where $a, b, c > 0$. Hint: The process $h \mapsto \mathbb{Z}(\sigma h - \mu)$ is equal in distribution to the process $h \mapsto \sqrt{\sigma}\mathbb{Z}(g) - \mathbb{Z}(\mu)$, where $\sigma \geq 0$ and $\mu \in \mathbb{R}$. Apply the change of variable $h = (a/b)^{2/3}g - c/(2b)$ and note that the location of a maximum does not change by multiplication by a positive constant or a vertical shift.

## 14.7 Notes

Theorem 14.1 and Lemma 14.2 are Theorem 3.2.2 and Lemma 3.2.1, respectively, of VW, while Theorem 14.4 and Corollary 14.5 are modified versions of Theorem 3.2.5 and Corollary 3.2.6 of VW. The monotone density estimation example in Section 14.5.2 is a variation of Example 3.2.14 of VW. The limiting behavior of the Grenander estimator of this example was obtained by Prakasa Rao (1969). Exercise 14.6.9 is an expanded version of Exercise 24.5 of van der Vaart (1998) and Exercise 14.6.10 is an expanded version of Exercise 3.2.5 of VW.

# 15
# Case Studies II

The examples of this section illustrate a variety of empirical process techniques applied in a statistical context. The first example is a continuation of the partly linear logistic regression example introduced in Chapter 1 and studied in some detail in Chapter 4. The issues addressed are somewhat technical, but they are also both instructive and necessary for securing the desired results. The second example utilizes empirical process results for independent but not identically distributed observations from Chapter 11 to address an issue in clinical trials. The third example applies Z-estimation theory for estimation and inference in the proportional odds regression model for right-censored survival data, while the fourth example considers hypothesis testing for the presence of a change-point in the regression model studied in Section 14.5.1. An interesting feature of this fourth example is that the model is partially unidentifiable under the null hypothesis of no change-point. The fifth example utilizes maximal inequalities (see Section 8.1) to establish asymptotic results for very high dimensional data sets which arise in gene microarray studies. These varied examples demonstrate how the empirical process methods of the previous chapters can be used to solve challenging and important problems in statistical inference.

## 15.1 Partly Linear Logistic Regression Revisited

As promised in Section 4.5, we now tie up some of the technical loose ends for establishing that the proposed penalized log-likelihood estimator $\hat{\beta}_n$ is

asymptotically efficient. Specifically, we establish (i) that $\mathcal{H}_c$ is Donsker for every $c < \infty$, (ii) that $J(\hat{\eta}_n) = O_P(1)$, (iii) that Expression (4.12) holds, and (iv) that both $\hat{\beta}_n$ and $\hat{\eta}_n$ are uniformly consistent. Along the way, we will also verify that the $L_2$ convergence of $\hat{\eta}_n$ occurs at the optimal rate $n^{-k/(2k+1)}$ (see Cox and O'Sullivan, 1990, and Mammen and van de Geer, 1997). We need to strengthen our previous assumptions to require the existence of a known $c_0 < \infty$ such that both $|\beta| < c_0$ and $\|\eta\|_\infty < c_0$ and also that the density of $U$ is strictly positive and finite.

For (i), Theorem 9.21 readily implies, after some rescaling of $\mathcal{H}_c$, that

$$\log N_{[]}(\epsilon, \mathcal{H}_c, L_2(P)) \quad \leq \quad M\epsilon^{-1/k}, \tag{15.1}$$

for all $\epsilon > 0$, where $M$ only depends on $c$ and $P$. This readily yields that $\mathcal{H}_c$ is Donsker.

To establish (ii), first define the composite parameter $\theta \equiv (\beta, \eta)$, and let $\ell_\theta(x) \equiv y w_\theta(x) - \log\left(1 + e^{w_\theta(x)}\right)$ and $w_\theta(x) \equiv z\beta + \eta(u)$. The quite technical Theorem 15.1 below yields that both $J(\hat{\eta}_n) = O_P(1)$ and

$$\|w_{\hat{\theta}_n}(X) - w_{\theta_0}(X)\|_{P,2} \quad = \quad O_P(n^{-k/(2k+1)}) \tag{15.2}$$

hold. Thus (4.12) holds trivially, and we are done with Parts (ii) and (iii).

We now use (15.2) to obtain (iv) and optimality of the $L_2$ convergence rate of $\hat{\eta}_n$. Recall that $\tilde{h}_1(u) \equiv E[Z|U = u]$ and that $P[Z - \tilde{h}_1(U)]^2 > 0$ by assumption. Since $E[(Z - \tilde{h}_1(U))g(U)] = 0$ for all $g \in L_2(U)$, (15.2) implies $|\hat{\beta}_n - \beta_0| = O_P(n^{-k/(2k+1)})$ and thus $\hat{\beta}_n$ is consistent. These results now imply that $P[\hat{\eta}_n(U) - \eta_0(U)]^2 = O_P(n^{-2k/(2k+1)})$, and thus we have $L_2$ optimality of $\hat{\eta}_n$ because of the assumptions on the density of $U$. Uniform consistency of $\hat{\eta}_n$ now follows form the fact that $J(\hat{\eta}_n) = O_P(1)$ forces $u \mapsto \hat{\eta}_n(u)$ to be uniformly equicontinuous in probability.

We now present Theorem 15.1:

THEOREM 15.1 *Under the given assumptions,* $J(\hat{\eta}_n) = O_P(1)$ *and*

$$\|w_{\hat{\theta}_n}(X) - w_{\theta_0}(X)\|_{P,2} = O_P(n^{-k/(2k+1)}).$$

**Proof.** We first need to establish a fairly precise bound on the bracketing entropy of

$$\mathcal{G} \equiv \left\{ \frac{\ell_\theta(X) - \ell_{\theta_0}(X)}{1 + J(\eta)} : |\beta - \beta_0| \leq c_1, \|\eta - \eta_0\|_\infty \leq c_1, J(\eta) < \infty \right\},$$

which satisfies $\mathcal{G} \subset \mathcal{G}_1 + \mathcal{G}_2(\mathcal{G}_1)$, where

$$\mathcal{G}_1 = \left\{ \frac{w_\theta(X) - w_{\theta_0}(X)}{1 + J(\eta)} : |\beta - \beta_0| \leq c_1, \|\eta - \eta_0\|_\infty \leq c_1, J(\eta) < \infty \right\},$$

$\mathcal{G}_2$ consists of all functions $t \mapsto a^{-1}\log(1 + e^{at})$ with $a \geq 1$, and $\mathcal{G}_2(\mathcal{G}_1) \equiv \{g_2(g_1) : g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2\}$. By (15.1) combined with the properties of

bracketing entropy and the fact that $J(\eta_0)$ is a finite constant, it is not hard to verify that there exists an $M_0 < \infty$ such that $\log N_{[]}(\epsilon, \mathcal{G}_1, L_2(P)) \leq M_0 \epsilon^{-1/k}$. Now by Exercise 15.6.1 combined with Lemma 15.2 below, we obtain that there exists a $K_1 < \infty$ such that $\log N_{[]}(\epsilon, \mathcal{G}_2(\mathcal{G}_1), L_2(P)) \leq K_1 \epsilon^{-1/k}$, for every $\epsilon > 0$. Combining this with preservation properties of bracketing entropy (see Lemma 9.25), we obtain that there exists an $M_1 < \infty$ such that $\log N_{[]}(\epsilon, \mathcal{G}, L_2(P)) \leq M_1 \epsilon^{-1/k}$ for all $\epsilon > 0$.

Combining this result with Theorem 15.3 below, we obtain that

$$\begin{aligned} (15.3) \quad \left|(\mathbb{P}_n - P)(\ell_{\hat{\theta}_n}(X) - \ell_{\theta_0}(X))\right| &= O_P(n^{-1/2}(1 + J(\hat{\eta}))) \\ &\times \left[\left\|\frac{\ell_{\hat{\theta}_n}(X) - \ell_{\theta_0}(X)}{1 + J(\hat{\eta}_n)}\right\|_{P,2}^{1-1/(2k)} \vee n^{-(2k-1)/[2(2k+1)]}\right]. \end{aligned}$$

Now note that by a simple Taylor expansion and the boundedness constraints on the parameters, there exists a $c_1 > 0$ and a $c_2 < \infty$ such that

$$P(\ell_{\hat{\theta}_n}(X) - \ell_{\theta_0}(X)) \leq -c_1 P[w_{\hat{\theta}_n}(X) - w_{\theta_0}(X)]^2$$

and

$$\left|\ell_{\hat{\theta}_n}(X) - \ell_{\theta_0}(X)\right| \leq c_2 \left|w_{\hat{\theta}_n}(X) - w_{\theta_0}(X)\right|,$$

almost surely. Combining this with (15.3) and a simple Taylor expansion, we can readily establish that

$$\begin{aligned} \lambda_n^2 J^2(\hat{\eta}_n) &\leq \lambda_n^2 J^2(\eta_0) + (\mathbb{P}_n - P)(\ell_{\hat{\theta}_n}(X) - \ell_{\theta_0}(X)) \\ &\quad + P\left[\ell_{\hat{\theta}_n}(X) - \ell_{\theta_0}(X)\right] \\ &\leq O_P(\lambda_n^2) + O_P(n^{-1/2})(1 + J(\hat{\eta}_n)) \\ &\quad \times \left[\left\|\frac{w_{\hat{\theta}_n}(X) - w_{\theta_0}(X)}{1 + J(\hat{\eta}_n)}\right\|_{P,2}^{1-1/(2k)} \vee n^{-(2k-1)/[2(2k+1)]}\right] \\ &\quad - c_1 P\left[w_{\hat{\theta}_n}(X) - w_{\theta_0}(X)\right]^2, \end{aligned}$$

from which we can deduce that both

$$\begin{aligned} (15.4) \quad \frac{J^2(\hat{\eta}_n)}{1 + J(\hat{\eta}_n)} &= O_P(1) + O_P\left(n^{(2k-1)/[2(2k+1)]}\right) \\ &\times \left[\left\|\frac{w_{\hat{\theta}_n}(X) - w_{\theta_0}(X)}{1 + J(\hat{\eta}_n)}\right\|_{P,2}^{1-1/(2k)} \vee n^{-(2k-1)/[2(2k+1)]}\right] \end{aligned}$$

and

$$\begin{aligned} (15.5) \quad \left\|\frac{w_{\hat{\theta}_n}(X) - w_{\theta_0}(X)}{1 + J(\hat{\eta}_n)}\right\|_{P,2}^2 &= O_P(\lambda_n^2) + O_P(n^{-1/2}) \\ &\times \left[\left\|\frac{w_{\hat{\theta}_n}(X) - w_{\theta_0}(X)}{1 + J(\hat{\eta}_n)}\right\|_{P,2}^{1-1/(2k)} \vee n^{-(2k-1)/[2(2k+1)]}\right]. \end{aligned}$$

Letting $A_n \equiv n^{k/(2k+1)}(1+J(\hat{\eta}_n))^{-1}\|w_{\hat{\theta}_n}(X) - w_{\theta_0}(X)\|_{P,2}$, we obtain from (15.5) that $A_n^2 = O_P(1) + O_P(1)A_n^{1-1/(2k)}$. Solving this yields $A_n = O_P(1)$, which implies

$$\frac{\|w_{\hat{\theta}_n}(X) - w_{\theta_0}(X)\|_{P,2}}{1+J(\hat{\eta}_n)} = O_P(n^{-k/(2k+1)}).$$

Applying this to (15.4) now yields $J(\hat{\eta}_n) = O_P(1)$, which implies $\|w_{\hat{\theta}_n}(X) - w_{\theta_0}(X)\|_{P,2} = O_P(n^{-k/(2k+1)})$, and the proof is complete.□

LEMMA 15.2 *For a probability measure $P$, let $\mathcal{F}_1$ be a class of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$, and let $\mathcal{F}_2$ denote a class of nondecreasing functions $f_2 : \mathbb{R} \mapsto [0,1]$ that are measurable for every probability measure. Then,*

$$\begin{aligned}\log N_{[]}(\epsilon, \mathcal{F}_2(\mathcal{F}_1), L_2(P)) \quad &\leq \quad 2\log N_{[]}(\epsilon/3, \mathcal{F}_1, L_2(P)) \\ &\quad + \sup_Q \log N_{[]}(\epsilon/3, \mathcal{F}_2, L_2(Q)),\end{aligned}$$

*for all $\epsilon > 0$, where the supremum is over all probability measures $Q$.*

**Proof.** Fix $\epsilon > 0$, and let $\{[f_k, g_k)], 1 \leq k \leq n_1\}$ be a minimal $L_2(P)$ bracketing $\epsilon/3$-cover for $\mathcal{F}_1$, where $f_k$ is the lower- and $g_k$ is the upper-boundary function for the bracket. For each $f_k$, construct a minimal $L_2(Q_{k,1})$ bracketing $\epsilon/3$-cover for $\mathcal{F}_1(f_k(x))$, where $Q_{k,1}$ is the distribution of $f_k(X)$. Let $n_2 = \sup_Q \log N_{[]}(\epsilon/3, \mathcal{F}_2, L_2(Q))$, and choose a corresponding minimal cover $\{[f_{k,j,1}, g_{k,j,1}], l \leq j \leq n_2\}$. Construct a similar cover $\{[f_{k,j,2}, g_{k,j,2}], 1 \leq j \leq n_2\}$ for each $\mathcal{F}_1(g_k(x))$, $1 \leq k \leq n_1$.

Let $h_1 \in \mathcal{F}_1$ and $h_2 \in \mathcal{F}_2$ be arbitrary; let $[f_k, g_k]$ be the bracket containing $h_1$; let $[f_{k,j,1}, g_{k,j,1}]$ be the bracket containing $h_2(f_k)$; and let $[f_{k,j,2}, g_{k,j,2}]$ be the bracket containing $h_2(g_k)$. Then $[f_{k,j,1}(f_k), g_{k,j,2}(g_k)]$ is an $L_2(P)$ $\epsilon$-bracket which satisfies $f_{k,j,1}(f_k) \leq h_2(f_k) \leq h_2(h_1) \leq h_2(g_k) \leq g_{k,j,2}(g_k)$. Thus, since $f_1$ and $f_2$ were arbitrary, the number of $L_2(P)$ $\epsilon$-brackets needed to completely cover $\mathcal{F}_2(\mathcal{F}_1)$ is bounded by

$$N_{[]}^2(\epsilon/3, \mathcal{F}_1, L_2(P)) \times \sup_Q N_{[]}(\epsilon/3, \mathcal{F}_2, L_2(Q)),$$

and the desired result follows.□

THEOREM 15.3 *Let $\mathcal{F}$ be a uniformly bounded class of measurable functions such that for some measurable $f_0$, $\sup_{f\in\mathcal{F}} \|f - f_0\|_\infty < \infty$. Moreover, assume that $\log N_{[]}(\epsilon, \mathcal{F}, L_2(P)) \leq K_1\epsilon^{-\alpha}$ for some $K_0 < \infty$ and $\alpha \in (0,2)$ and for all $\epsilon > 0$. Then*

$$\sup_{f\in\mathcal{F}} \left[ \frac{|(\mathbb{P}_n - P)(f - f_0)|}{\|f - f_0\|_{P,2}^{1-\alpha/2} \vee n^{-(2-\alpha)/[2(2+\alpha)]}} \right] = O_P(n^{-1/2}).$$

**Proof.** This is a result presented on Page 79 of van de Geer (2000) and follows from Lemma 5.13 on the same page, the proof of which can be found in Pages 79–80.□

## 15.2 The Two-Parameter Cox Score Process

In this section, we apply the results of Section 11.4 to study the two-parameter Cox score process in a sequential clinical trial with staggered entry of patients. The material of this section comes from Section 4 of Kosorok (2003). We will make the somewhat weak assumption that patients are independent but not necessarily i.i.d. Such a generalization is necessary if an adaptive design, such as a biased coin design (see Wei, 1978), is used rather than a simple i.i.d. randomized design. The two time parameters are the calendar time of entry and the time since entry for each patient. Sellke and Siegmund (1983) made a fundamental breakthrough on this problem and important work on this problem has also been done by Slud (1984) and Gu and Lai (1991). The most recent work on this problem, and the work upon which we will build, is that of Bilias, Gu, and Ying (1997) (hereafter abbreviated BGY).

For each patient, there is an entry time ($\tau_i \geq 0$), a continuous failure time ($X_i \geq 0$), a censoring time ($C_i \geq 0$), and a real valued covariate process $Z_i = \{Z_i(s), s \geq 0\}$, $1 \leq i \leq n$, $n \geq 1$. Although the results we present are valid when $Z_i$ is vector valued, we will assume that $Z_i$ is scalar valued for simplicity. The entry time is on the calendar time scale while the remaining quantities are on the time since entry time scale. As is done in BGY, we will assume that the quadruples $(\tau_i, X_i, C_i, Z_i)$, $i = 1 \ldots n$, are independent (but not identically distributed) and that the conditional hazard rate of $X_i$ at $s$, give $\tau_i$, $C_i$, and $Z_i(u)$, for $u \leq s$, has the Cox proportional hazards form $\exp[\beta Z_i(s)]\,\lambda_0(s)$, for some unknown baseline hazard $\lambda_0$. Thus, at calendar time $x$, the $i$th individual's failure time is censored at $C_i \wedge (x - \tau_i)^+$, where for a real number $u$, $u^+ = u\{u \geq 0\}$. At any calendar time $x$, what we actually observe under possibly right-censoring is $U_i(x) = X_i \wedge C_i \wedge (x - \tau_i)^+$ and $\Delta_i(x) = \{X_i \leq C_i \wedge (x - \tau_i)^+\}$. For $x \geq s$, denote $N_i(x, s) = \Delta_i(x)\,\{U_i(x) \leq s\}$, $Y_i(x, s) = \{U_i(x) \geq s\}$, and

$$\overline{Z}(\beta; x, s) = \frac{\sum_{i=1}^n Z_i(s) \exp[\beta Z_i(s)]\, Y_i(x, s)}{\sum_{i=1}^n \exp[\beta Z_i(s)]\, Y_i(x, s)}.$$

The statistic of interest we will focus on is

$$W_n(x, s) = n^{-1/2} \sum_{i=1}^n \int_0^s \left[ Z_i(u) - \overline{Z}(\beta_0; x, s) \right] N_i(x, du)$$

under the null hypothesis that $\beta_0$ is the true value of $\beta$ in the foregoing Cox proportional hazards model. For ease of exposition, we will assume throughout that $\beta_0$ is known. It is easy to show that $W_n(x, x)$ is the partial likelihood score at calendar time $x$ and at $\beta = \beta_0$. When $Z_i$ is a dichotomous treatment indicator, $W_n(x, x)$ is the well known two-sample log-rank statistic and is a good choice for testing $H_0 : \beta = \beta_0$

versus $H_A : \beta \neq \beta_0$. However, for certain other alternative hypotheses, the supremum version of $W_n$, $\sup_{s\in(0,x]} |W_n(x,s)|$, is a more powerful statistic (see Fleming, Harrington, and O'Sullivan, 1987). Determining the asymptotic behavior of this last statistic requires weak convergence results whether continuous or group sequential interim monitoring is being done. Let $D(x_0) = \{(x,s) : 0 \le s \le x \le x_0\}$, where $x_0$ satisfies some assumptions given below. Under several assumptions, which we will review later, BGY demonstrate that $W_n(x,s)$ converges weakly in the uniform topology on $\ell^\infty(D(x_0))$ to a tight mean zero Gaussian process $W$ with a covariance which we will denote $H$.

Unfortunately, the complexity of the distribution of $W$ precludes it from being used directly to compute critical boundaries, especially for the supremum version of the test statistic, during clinical trial execution. We will now show that the results of Section 11.4 can be used to obtain an asymptotically valid Monte Carlo estimate of the distribution of $W$ for this purpose, something like a "wild bootstrap." We now present the assumptions given by BGY:

(a) $x_0 < \infty$ satisfies $\liminf_{n\to\infty} n^{-1} \sum_{i=1}^n \mathrm{E} Y_i(x_0, x_0) > 0$;

(b) (Condition 1 in BGY) There exists a nonrandom $B < \infty$ such that the total variation $|Z_i(0)| + \int_0^{x_0} |Z_i(du)| \le B$;

(c) (Condition 2 in BGY) For $k = 0, 1, 2$, there exists $\Gamma_k(x,s)$ such that, for all $(x,s) \in D(x_0)$,

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^n \mathrm{E}\left[Z_i^k(s) Y_i(x,s) \exp(\beta_0 Z_i(s))\right] = \Gamma_k(x,s);$$

(d) (Condition 3 in BGY) Letting $K(x,s) = \Gamma_1(x,s)/\Gamma_0(x,s)$ and

$$K_n(x,s) = \frac{\sum_{i=1}^n \mathrm{E}\left[Z_i(s) Y_i(x,s) \exp(\beta_0 Z_i(s))\right]}{\sum_{i=1}^n \mathrm{E}\left[Y_i(x,s) \exp(\beta_0 Z_i(s))\right]},$$

then, for each fixed $s$, $K(\cdot, s)$ is continuous on $[s, x_0]$ and

$$\lim_{n\to\infty} \sup_{0\le x\le x_0} \int_0^x \left[K_n(x,s) - K(x,s)\right]^2 ds = 0.$$

In the proof of their Theorem 2.2, BGY establish that $W_n(x,s)$ converges in probability to $\check{W}_n(x,s) = \sum_{i=1}^n f_{ni}(\omega, (x,s))$, where

$$f_{ni}(\omega,(x,s)) \quad = \quad n^{-1/2} \int_0^s \left[Z_i(u) - K_n(x,u)\right] M_i(x, du), \tag{15.6}$$

$M_i(x,s) = N_i(x,s) - \int_0^s Y_i(x,u) \exp(\beta_0 Z_i(u)) \lambda_0(u) du$, $\omega \in \Omega$, and where the data for this statistic comes from the probability space $\{\Omega, \mathcal{W}, \Pi\}$. BGY

establish that all of the conditions of Pollard's (1990) functional central limit theorem are satisfied with measurable envelope functions, hence all of the conditions of Theorem 11.16 are satisfied, except for the "sufficient measurability" condition. It is actually not clear how all the necessary measurability conditions are addressed by BGY; however, it is not difficult to show that the inherent double right-continuity of both $\breve{W}_n(x,s)$ and all of the array components $f_{ni}(\omega,(x,s))$ gives us separability (which implies AMS by Lemma 11.15) with $T = D(x_0)$ and $T_n = D(x_0) \cap [\mathcal{Q} \cup \{x_0\}]^2$, where $\mathcal{Q}$ is the set of rationals.

All of the conditions of Theorem 11.16 are thus satisfied and $W_n$ converges weakly to a tight, mean zero Gaussian process $W$. Let $\{z_i, i \geq 1\}$ be a random sequence satisfying Condition (F) of Section 11.4.2 and which is independent of the data generating $W_n$. Since we usually do not have exact knowledge of either $K_n$ or the $M_i$s, we will need to find appropriate estimators and then apply Theorem 11.18 to obtain a Monte Carlo method of estimating the distribution of $W$ based only on the data from the clinical trial. In this setting, $\mu_{ni}((x,s)) = 0$. However, if we use the estimator

$$\begin{aligned}(15.7)\ \hat{\mu}_{ni}(\omega,(x,s)) &= n^{-1/2}\int_0^s \left[\overline{Z}(\beta_0;x,u) - K_n(x,u)\right] M_i(x,du) \\ &\quad -n^{-1/2}\int_0^s \left[Z_i(u) - \overline{Z}(\beta_0;x,s)\right] \hat{M}_i(x,du),\end{aligned}$$

where

$$\hat{M}_i(x,du) = M_i(x,du) - Y_i(x,u)\exp(\beta_0 Z_i(u))\frac{\sum_{j=1}^n M_j(x,du)}{\sum_{j=1}^n Y_j(x,u)\exp(\beta_0 Z_j(u))},$$

then

$$\begin{aligned}&f_{ni}(\omega,(x,s)) - \hat{\mu}_{ni}(\omega,(x,s)) \\ &\quad = n^{-1/2}\int_0^s \left[Z_i(u) - \overline{Z}(\beta_0;x,u)\right] \hat{M}_i(x,du) \\ &\quad = n^{-1/2}\int_0^s \left[Z_i(u) - \overline{Z}(\beta_0;x,s)\right] \\ &\qquad \times \left[N_i(x,du) - Y_i(x,u)\exp(\beta_0 Z_i(u))\frac{\sum_{j=1}^n N_j(x,du)}{\sum_{j=1}^n Y_j(x,u)\exp(\beta_0 Z_j(u))}\right]\end{aligned}$$

can be computed from the data.

If we can next establish that the array $\{\hat{\mu}_{ni}\}$ satisfies the conditions of Theorem 11.18, we will then have a Monte Carlo method of estimating the distribution of $W$ based on the conditional distribution of the process

$$\hat{W}_n(x,s) = \sum_{i=1}^n z_i \left[f_{ni}(\omega,(x,s)) - \hat{\mu}_{ni}(\omega,(x,s))\right]$$

given the data from the clinical trial. To this end, we have the following Theorem:

THEOREM 15.4 *In the two-parameter Cox score process setting with staggered entry of patients with Assumptions (a) through (d) satisfied, the triangular array of estimators* $\{\hat{\mu}_{ni}\}$*—of the form given in (15.7)—satisfies Conditions (G) through (I) of Theorem 11.18.*

**Proof.** Because the processes $\hat{\mu}_{ni}$ possess the same double right continuity that $W_n$ possesses, the separability—and hence AMS—condition is readily satisfied. Because each $\hat{\mu}_{ni}$ can be written as a difference between a monotone increasing and a monotone decreasing part with envelope $\hat{F}_{ni}(\omega)$, we have manageability since sums of monotone processes have pseudodimension one (see Lemma A.2 of BGY) and since sums of manageable processes are manageable. Since Assumption (a) implies $\Lambda_0(x_0) \equiv \int_0^{x_0} \lambda_0(u)du < \infty$ and also because of the bounded total variation of $Z_i$, we can use envelopes $\hat{F}_{ni}(\omega) = \hat{C}n^{-1/2}$, where there exists a $k < \infty$ not depending on $n$ such that $\hat{C} \vee k$ converges in probability to $k$ as $n \to \infty$. Hence Conditions (H) and (I) of Theorem 11.18 are satisfied. All that remains to be shown is that $\sup_{(t,s)\in T} \sum_{i=1}^n \hat{\mu}_{ni}^2(\omega,(x,s))$ converges to zero in probability.

Clearly,

$$\begin{aligned}(15.8)\quad &\sup_{(x,s)\in T} \sum_{i=1}^n \hat{\mu}_{ni}^2(\omega,(x,s)) \\ &\qquad \leq \quad 2 \sup_{(x,s)\in T} n^{-1} \sum_{i=1}^n \left( \int_0^x \left[\overline{Z}(\beta_0;x,u) - K_n(x,u)\right] M_i(x,du) \right)^2 \\ &+ \sup_{(x,s)\in T} \frac{2}{n} \sum_{i=1}^n \left( \int_0^x \left[Z_i(u) - \overline{Z}(\beta_0;x,u)\right] Y_i(x,u) \exp(\beta Z_i(u)) \overline{M}(x,du) \right)^2,\end{aligned}$$

where

$$\overline{M}(x,du) \equiv \left[ \sum_{i=1}^n Y_i(x,u) \exp(\beta_0 Z_i(u)) \right]^{-1} \left[ \sum_{i=1}^n M_i(x,du) \right].$$

However, since $\Lambda_0(x_0) < \infty$ and since $M_i$ is the difference between two nonnegative monotone functions bounded by $\Lambda_0(x_0)$, the first term on the right-hand-side of (15.8) is bounded by $k_1 \sup_{(x,s)\in T} \left|\overline{Z}(\beta_0;x,s) - K_n(x,s)\right|^2$, for some $k_1 < \infty$, thus this term vanishes in probability as $n \to \infty$. Using integration by parts combined with the fact that the total variation of $Z_i$ is bounded, it can be shown that there exists constants $c_1 < \infty$ and $c_2 < \infty$ such that the second term on the right-hand-side of (15.8) is bounded by

$$\sup_{(x,s)\in T} n^{-1}\sum_{i=1}^{n}\left(2c_1 \sup_{u\in[0,s]}\left|\overline{M}(x,u)\right| + 2c_2 \sup_{u\in[0,s]}\left|\int_0^u \overline{Z}(\beta_0;x,v)\overline{M}(x,dv)\right|\right)^2$$
$$\leq \quad 4c_1^2 \sup_{(x,s)\in T}\left|\overline{M}(x,s)\right|^2 + 4c_2 \sup_{(x,s)\in T}\left|\int_0^s \overline{Z}(\beta_0;x,u)\overline{M}(x,du)\right|^2;$$

and both of these terms can be shown to converge to zero in probability, as $n \to \infty$, by using arguments contained in BGY.□

## 15.3 The Proportional Odds Model Under Right Censoring

In the right-censored regression set-up, we observe $X = (U, \delta, Z)$, where $U = T \wedge C$, $\delta = 1\{U = T\}$, $Z \in \mathbb{R}^d$ is a covariate vector, $T$ is a failure time of interest, and $C$ is a right censoring time. We assume that $C$ and $T$ are independent given $Z$. The proportional odds regression model stipulates that the survival function of $T$ given $Z$ has the form

$$S_Z(t) \quad = \quad \left(1 + e^{\beta' Z} A(t)\right)^{-1}, \tag{15.9}$$

where $t \mapsto A(t)$ is nondecreasing on $[0, \tau]$, with $A(0) = 0$ and $\tau < \infty$ being the upper limit of the censoring distribution, i.e., we assume that $\mathrm{P}(C > \tau) = 0$. We also assume that $\mathrm{P}(C = \tau) > 0$, that var$[Z]$ is positive definite, and that the distribution of $Z$ and $C$ are uninformative of $S_Z$.

Let the true parameter values be denoted $\beta_0$ and $A_0$. We make the additional assumptions that the support of $Z$ is compact and that $\beta_0$ lies in a known compact $\subset \mathbb{R}^d$. Murphy, Rossini and van der Vaart (1997) develop asymptotic theory for maximum likelihood estimation of this model under general conditions for $A_0$ which permit ties in the failure time distribution. To simplify the exposition, we will make stronger assumptions on $A_0$ in order to facilitate arguments similar to those used in Lee (2000) and Kosorok, Lee and Fine (2004). Specifically, we assume that $A_0$ has a derivative $a_0$ that satisfies $0 < \inf_{t\in[0,\tau]} a_0(t) \leq \sup_{t\in[0,\tau]} a_0(t) < \infty$.

Let $F_Z \equiv 1 - S_Z$. For distinct covariate values $Z_1$ and $Z_2$, we can deduce from (15.9) that

$$\frac{F_{Z_1}(t)S_{Z_2}(t)}{F_{Z_2}(t)S_{Z_1}(t)} = \frac{e^{\beta' Z_1}}{e^{\beta' Z_2}},$$

which justifies the "proportional odds" designation. A motivation for this model is that in some settings it can be easier to justify on scientific grounds than other common alternatives such as the proportional hazards or accelerated failure time models (Murphy, Rossini and van der Vaart, 1997).

Define the composite model parameter $\theta \equiv (\beta, A)$. In the following sections, we derive the nonparametric maximum likelihood estimator

(NPMLE) $\hat{\theta}_n$, prove its existence, establish consistency and weak convergence, and verify that the bootstrap procedure is valid for all model parameters. Certain score and information operators will be needed for the weak convergence component, and these will be introduced just before we establish weak convergence but after we have proven consistency. While $\beta$ is assumed to lie in a known compact set, we make no such restrictions on $A$ for estimation. Hence the NPMLE $\hat{A}_n$ might be unbounded: for this reason, we need to verify that $\hat{A}_n(\tau) = O_P(1)$ and thus $\hat{\theta}_n$ "exists."

### *15.3.1 Nonparametric Maximum Likelihood Estimation*

The likelihood for a sample of $n$ i.i.d. observations $(U_1, \delta_1, Z_1), \ldots, (U_n, \delta_n, Z_n)$ is

$$\ell_n(\theta) = \mathbb{P}_n \left\{ \delta(\log a(U)) + \beta' Z) - (1+\delta) \log \left( 1 + e^{\beta' Z} A(U) \right) \right\},$$

where $a$ is the derivative of $A$. As discussed in Murphy (1994), maximizing $\ell_n$ over $A$ for fixed $\beta$ results in an maximizer that is piecewise constant with jumps at the observed failure times and thus does not have a continuous density. To address this issue, Murphy (1994) and Parner (1998) suggest replacing $a(u)$ in $\ell_n$ with $n\Delta A(u)$, the jump size of $A$ at $u$, which modified "empirical log-likelihood" we will denote $L_n(\theta)$. We will show later that when $A_0$ is continuous, the step sizes of the maximizer over $A$ of $L_n$ will go to zero as $n \to \infty$.

The procedure we will use to estimate $\theta$ is to first maximize the profile log-likelihood $pL_n(\beta) \equiv \sup_A L_n(\beta, A)$ to obtain $\hat{\theta}_n$. The associated maximizer over $A$ we will denote $\hat{A}_n$. In other words, $\hat{A}_n = \hat{A}_{\hat{\beta}_n}$, where $\hat{A}_\beta \equiv \text{argmax}_A L_n(\beta, A)$. We also define $\hat{\theta}_\beta \equiv (\beta, \hat{A}_\beta)$. Obviously $\hat{\theta}_n \equiv \hat{\theta}_{\hat{\beta}_n}$ is just the joint maximizer of $L_n(\theta)$. To characterize $\hat{A}_\beta$, consider one-dimensional submodels for $A$ defined by the map

$$t \mapsto A_t \equiv \int_0^{(\cdot)} (1 + t h_1(s)) dA(s),$$

where $h_1$ is an arbitrary total variation bounded cadlag function on $[0, \tau]$. The derivative of $L_n(\theta, A_t)$ with respect to $t$ evaluated at $t = 0$ is the score function for $A$:

(15.10)

$$V^\tau_{n,2}(\theta)(h_1) \equiv \mathbb{P}_n \left\{ \int_0^\tau h_1(s) dN(s) - (1+\delta) \left[ \frac{e^{\beta' Z} \int_0^{U \wedge \tau} h_1(s) dA(s)}{1 + e^{\beta' Z} A(U \wedge \tau)} \right] \right\},$$

where the subscript "2" denotes that this is the score of the second parameter $A$. The dependence on $\tau$ will prove useful in later sections, but,

for now, it can be ignored since $P(U \leq \tau) = 1$ by assumption. Choose $h_1(u) = 1\{u \leq t\}$, insert into (15.10), and equate the result to zero to obtain

$$\mathbb{P}_n N(t) = \mathbb{P}_n \left\{ \frac{(1+\delta)e^{\beta' Z} \int_0^t Y(s) d\hat{A}_\beta(s)}{1 + e^{\beta' Z} \hat{A}_\beta(U)} \right\}, \tag{15.11}$$

where $N(t) \equiv \delta 1\{U \leq t\}$ and $Y(t) \equiv 1\{U \geq t\}$ are the usual counting and at-risk processes for right-censored survival data.

Next define

$$W(t;\theta) \equiv \frac{(1+\delta)e^{\beta' Z} Y(t)}{1 + e^{\beta' Z} A(U)}$$

and solve (15.11) to obtain

$$\hat{A}_\beta(t) = \int_0^t \left\{ \mathbb{P}_n W(s; \hat{\theta}_\beta) \right\}^{-1} \mathbb{P}_n dN(s). \tag{15.12}$$

Thus $\hat{A}_\beta$ can be characterized as a stationary point of (15.12). This structure will prove useful in later developments and can also be used to calculate $\hat{\theta}_n$ from data. One approach to accomplishing this is by first using (15.12) to facilitate calculating $pL_n(\beta)$ so that $\hat{\beta}_n$ can be determined via a simple search algorithm and then taking $\hat{A}_n$ to be the solution of (15.12) corresponding to $\beta = \hat{\beta}_n$. In the case of multiple solutions, we take the one corresponding to the maximizer of $A \mapsto L_n(\hat{\beta}_n, A)$.

### *15.3.2 Existence*

While we are assuming that $\beta$ lies in a known, compact $\subset \mathbb{R}^d$, we are not setting boundedness restrictions on $A$. Fortunately, such restrictions are not necessary, as can be seen in the following lemma, which is the contribution of this section:

LEMMA 15.5 *Under the given conditions,* $\limsup_{n\to\infty} \hat{A}_n(\tau) < \infty$ *with inner probability one.*

**Proof.** Note that

$$\|\mathbb{P}_n N - Q_0\|_\infty \overset{\text{as*}}{\to} 0 \quad \text{and} \quad \sup_{u \geq 0} |(\mathbb{P}_n - P)1\{U \leq u\}| \overset{\text{as*}}{\to} 0, \tag{15.13}$$

where $Q_0 \equiv PN$. Let $X_1, X_2, \ldots$ be a fixed data sequence satisfying (15.13). Note that, without loss of generality, all data sequences will satisfy this since the probability of such a sequence is 1 by definition of outer-almost sure convergence.

Under this set-up, we can treat the maximum likelihood estimators as a sequence of fixed quantities. By the assumed compactness of the parameter

space for $\beta$, there exists a subsequence of $\{n\}$ for which $\hat{\beta}_n$ converges to a bounded vector $\beta_*$ along that subsequence. Now choose a further subsequence $\{n_k\}$ for which both $\hat{\beta}_{n_k} \to \beta_* \in \mathbb{R}^d$ and $\hat{A}_{n_k}(\tau) \to \infty$. We will now work towards a contradiction.

Let $\theta_n \equiv (\beta_0, \mathbb{P}_n N)$, and note that, by definition of the NPMLE,

$$\begin{aligned} 0 &\leq L_{n_k}(\hat{\theta}_n) - L_{n_k}(\theta_n) \\ &\leq O(1) + \int_0^\tau \log(n\Delta A_{n_k}(s))\mathbb{P}_{n_k} dN(s) \\ &\quad -\mathbb{P}_{n_k}\left[(1+\delta)\log(1+\hat{A}_{n_k}(U))\right]. \end{aligned} \tag{15.14}$$

Let $\{u_0, u_1, \ldots, u_M\}$ be a partition of $[0, \tau]$ for some finite $M$, with $0 = u_0 < u_1 < \cdots < u_M = \tau$, which we will specify in more detail shortly, and define $N^j(s) \equiv N(s)1\{U \in [u_{j-1}, u_j]\}$ for $1 \leq j \leq M$. Now by Jensen's inequality,

$$\begin{aligned} \int_0^\tau \log(n_k \Delta \hat{A}_{n_k})\mathbb{P}_{n_k} dN^j(s) &\leq \mathbb{P}_{n_k} N^j(\tau) \log\left(\frac{\int_0^{u_j} n\Delta \hat{A}_{n_k}(s) d\mathbb{P}_n N^j(s)}{\mathbb{P}_{n_k} N^j(\tau)}\right) \\ &\leq O(1) + \log(\hat{A}_{n_k}(u_j))\mathbb{P}_{n_k}\delta 1\{U \in [u_{j-1}, u_j]\}. \end{aligned}$$

Thus the right-side of (15.14) is dominated by

$$\begin{aligned} &O(1) + \sum_{j=1}^{M-1} \log \hat{A}_{n_k}(u_j) \\ &\quad \times \mathbb{P}_{n_k}\left(\delta 1\{U \in [u_{j-1}, u_j]\} - (1+\delta)1\{U \in [u_j, u_{j+1}]\}\right) \\ &\quad + \log \hat{A}_{n_k}(\tau)\mathbb{P}_{n_k}\left(\delta 1\{U \in [u_{M-1}, \infty]\} - (1+\delta)1\{U \in [\tau, \infty]\}\right). \end{aligned} \tag{15.15}$$

Without loss of generality, assume $Q_0(\tau) > 0$. Because of the assumptions, we can choose $u_0, u_1, \ldots, u_M$ with $M < \infty$ such that for some $\eta > 0$,

$$P\left((1+\delta)1\{U \in [\tau, \infty]\}\right) \geq \eta + P\left(\delta 1\{U \in [u_{M-1}, \infty]\}\right)$$

and

$$P\left((1+\delta)1\{U \in [u_j, u_{j+1}]\}\right) \geq \eta + P\left(\delta 1\{U1[u_{j-1}, u_j]\}\right),$$

for all $1 \leq j \leq M-1$. Hence

$$(15.15) \leq -(\eta + o(1)) \log \hat{A}_{n_k}(\tau) \to -\infty,$$

as $k \to \infty$, which yields the desired contradiction. Thus the conclusions of the lemma follow since the data sequence was arbitrary.□

### *15.3.3 Consistency*

In this section, we prove uniform consistency of $\hat{\theta}_n$. Let $\Theta \equiv \mathcal{B}_0 \times \mathcal{A}$ be the parameter space for $\theta$, where $\mathcal{B}_0 \subset \mathbb{R}^d$ is the known compact containing $\beta_0$ and $\mathcal{A}$ is the collection of all monotone increasing functions $A : [0, \tau] \mapsto [0, \infty]$ with $A(0) = 0$. The following is the main result of this section:

THEOREM 15.6 *Under the given conditions,* $\hat{\theta}_n \overset{\text{as*}}{\to} \theta_0$.

**Proof.** Define $\tilde{\theta}_n = (\beta_0, \tilde{A}_n)$, where $\tilde{A}_n \equiv \int_0^{(\cdot)} [PW(s;\theta_0)]^{-1} \mathbb{P}_n dN(s)$. Note that

(15.16)

$$\begin{aligned} L_n(\hat{\theta}_n) - L_n(\tilde{\theta}_n) &= \int_0^\tau \frac{PW(s;\theta_0)}{\mathbb{P}_n W(s;\hat{\theta}_n)} \mathbb{P}_n dN(s) + (\hat{\beta}_n - \beta_0)' \mathbb{P}_n \int_0^\tau Z dN(s) \\ &\quad - \mathbb{P}_n \left[ (1+\delta) \log \left( \frac{1 + e^{\hat{\beta}_n' Z} \hat{A}_n(U)}{1 + e^{\beta_0' Z} \tilde{A}_n(U)} \right) \right]. \end{aligned}$$

By Lemma 15.7 below, $(\mathbb{P}_n - P) W(t;\hat{\theta}_n) \overset{\text{as*}}{\to} 0$. Combining this with Lemma 15.5 yields that $\liminf_{n\to\infty} \inf_{t\in[0,\tau]} \mathbb{P}_n W(t;\hat{\theta_n}) > 0$ and that the $\limsup_{n\to\infty}$ of the total variation of $t \mapsto \left[ PW(t;\hat{\theta}_n) \right]^{-1}$ is $< \infty$ with inner probability one. Since $\left\{ \int_0^t g(s) dN(s) : t \in [0,\tau], g \in D[0,\tau], \text{the total variation of } g \le M \right\}$ is Donsker for every $M < \infty$, we now have

$$\int_0^\tau \frac{PW(s;\theta_0)}{\mathbb{P}_n W(s;\hat{\theta}_n)} \mathbb{P}_n dN(s) - \int_0^\tau \frac{PW(s;\theta_0)}{PW(s;\hat{\theta}_n)} dQ_0(s) \overset{\text{as*}}{\to} 0. \tag{15.17}$$

Combining Lemma 15.5 with the fact that

$$\left\{ (1+\delta) \log \left( 1 + e^{\beta' Z} A(U) \right) : \theta \in \Theta, A(\tau) \le M \right\}$$

is Glivenko-Cantelli for each $M < \infty$ yields

$$(\mathbb{P}_n - P) \left[ (1+\delta) \log \left( \frac{1 + e^{\hat{\beta}_n' Z} \hat{A}_n(U)}{1 + e^{\beta_0' Z} A_0(U)} \right) \right] \overset{\text{as*}}{\to} 0. \tag{15.18}$$

Now combining results (15.17) and (15.18) with (15.16), we obtain that

$$\begin{aligned} L_n(\hat{\theta}_n) - L_n(\tilde{\theta}_n) - \int_0^\tau \frac{PW(s;\theta_0)}{PW(s;\hat{\theta}_n)} dQ_0(s) - (\hat{\beta}_n - \beta_0)' P[Z\delta] \\ + P \left[ (1+\delta) \log \left( \frac{1 + e^{\hat{\beta}_n' Z} \hat{A}_n(U)}{1 + e^{\beta_0' Z} A_0(U)} \right) \right] \\ \overset{\text{as*}}{\to} 0. \end{aligned} \tag{15.19}$$

Now select a fixed sequence $X_1, X_2, \ldots$ for which the previous convergence results hold, and note that such sequences occur with inner probability one. Reapplying (15.12) yields

$$\limsup_{n\to\infty} \sup_{s,t\in[0,\tau]} \frac{|\hat{A}_n(s) - \hat{A}_n(t)|}{|\mathbb{P}_n(N(s) - N(t))|} < \infty.$$

Thus there exists a subsequence $\{n_k\}$ along which both $\|\hat{A}_{n_k} - \tilde{A}\|_\infty \to 0$ and $\hat{\beta}_{n_k} \to \tilde{\beta}$, for some $\tilde{\theta} = (\tilde{\beta}, \tilde{A})$, where $\tilde{A}$ is both continuous and bounded. Combining this with (15.19), we obtain

$$0 \le L_{n_k}(\hat{\theta}_{n_k}) - L_{n_k}(\tilde{\theta}_n) \to P_0 \log\left[\frac{dP_{\tilde{\theta}}}{dP_0}\right] \le 0,$$

where $P_\theta$ is the probability measure of a single observation on the specified model at parameter value $\theta$ and where $P_0 \equiv P_{\theta_0}$. Since the "model is identifiable" (see Exercise 15.6.2 below), we obtain that $\hat{\theta}_n \to \theta_0$ uniformly. Since the sequence $X_1, X_2, \ldots$ was an arbitrary representative from a set with inner probability one, we obtain that $\hat{\theta}_n \to \theta_0$ almost surely.

Since $\hat{A}_n$ is a piecewise constant function with jumps $\Delta\hat{A}_n$ only at observed failure times $t_1, \ldots, t_{m_n}$, $\hat{\theta}_n$ is a continuous function of a maximum taken over $m_n + d$ real variables. This structure implies that $\sup_{t\in[0,\tau]} |\hat{A}_n(t) - A_0(t)|$ is a measurable random variable, and hence the uniform distance between $\hat{\theta}_n$ and $\theta_0$ is also measurable. Thus the almost sure convergence can be strengthened to the desired outer almost sure convergence.□

LEMMA 15.7 *The class of functions* $\{W(t;\theta) : t \in [0,\tau], \theta \in \Theta\}$ *is P-Donsker.*

**Proof.** It is fairly easy to verify that

$$\mathcal{F}_1 \equiv \left\{(1+\delta)e^{\beta' Z} Y(t) : t \in [0,\tau], \beta \in \mathcal{B}_0\right\}$$

is a bounded $P$-Donsker class. If we can also verify that

$$\mathcal{F}_2 \equiv \left\{\left(1 + e^{\beta' Z} A(U)\right)^{-1} : \theta \in \Theta\right\}$$

is $P$-Donsker, then we are done since the product of two bounded Donsker classes is also Donsker. To this end, let $\phi : \mathbb{R}^2 \mapsto \mathbb{R}$ be defined by

$$\phi(x,y) = \frac{1-y}{1-y+e^x y},$$

and note that $\phi$ is Lipschitz continuous on sets of the form $[-k,k] \times [0,1]$, with a finite Lipschitz constant depending only on $k$, for all $k < \infty$ (see Exercise 15.6.3 below). Note also that

$$\mathcal{F}_2 = \left\{ \phi\left(\beta' Z, \frac{A(U)}{1+A(U)}\right) : \theta \in \Theta \right\}.$$

Clearly, $\{\beta' Z : \beta \in \mathcal{B}_0\}$ is Donsker with range contained in $[-k_0, k_0]$ for some $k_0 < \infty$ by the given conditions. Moreover, $\{A(U)(1+A(U))^{-1} : A \in \mathcal{A}\}$ is a subset of all monotone, increasing functions with range $[0,1]$ and thus, by Theorem 9.24, is Donsker. Hence, by Theorem 9.31, $\mathcal{F}_2$ is $P$-Donsker, and the desired conclusions follow.□

### *15.3.4 Score and Information Operators*

Because we have an infinite dimensional parameter $A$, we need to take care with score and information operator calculations. The overall idea is that we need these operators in order to utilize the general Z-estimator convergence theorem (Theorem 2.11) in the next section to establish asymptotic normality of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ and obtain validity of the bootstrap. To facilitate the development of these operators, let $\mathcal{H}$ denote the space of elements $h = (h_1, h_2)$ with $h_1 \in \mathbb{R}^d$ and $h_2 \in D[0,\tau]$ of bounded variation. We supply $\mathcal{H}$ with the norm $\|h\|_{\mathcal{H}} \equiv \|h_1\| + \|h_2\|_v$, where $\|\cdot\|$ is the Euclidean norm and $\|\cdot\|_v$ is the total variation norm.

Define $\mathcal{H}_p \equiv \{h \in \mathcal{H} : \|h\|_{\mathcal{H}} \leq p\}$, where the inequality is strict when $p = \infty$. The parameter $\theta$ can now be viewed as an element of $\ell^\infty(\mathcal{H}_p)$ if we define

$$\theta(h) \equiv h_1'\beta + \int_0^\tau h_2(s) dA(s), \quad h \in \mathcal{H}_p, \quad \theta \in \Theta.$$

Note that $\mathcal{H}_1$ is sufficiently rich to be able to extract out all components of $\theta$. For example, $\tilde{h} = (h_1, 0) = ((0,1,0,\ldots)', 0)$ extracts out the second component of $\beta$, i.e., $\beta_2 = \theta(\tilde{h})$, while $\tilde{h}_{*,u} = (0, 1\{(\cdot) \leq u\})$ extracts out $A(u)$, i.e., $A(u) = \theta(\tilde{h}_{*,u})$. As a result, the parameter space $\Theta$ becomes a subset of $\ell^\infty(\mathcal{H}_p)$ with norm $\|\theta\|_{(p)} \equiv \sup_{h \in \mathcal{H}_p} |\theta(h)|$. We can study weak convergence in the uniform topology of $\hat{\theta}_n$ via this functional representation of the parameter space since, for all $1 \leq p < \infty$ and every $\theta \in \Theta$, $\|\theta\|_\infty \leq \|\theta\|_{(p)} \leq 4p\|\theta\|_\infty$ (see Exercise 15.6.4).

We now calculate the score operator which will become the Z-estimating equation to which we will apply the Z-estimator convergence theorem. Consider the one-dimensional submodel defined by the map

$$t \mapsto \theta_t \equiv \theta + t(h_1, \int_0^{(\cdot)} h_2(s) dA(s)), \quad h \in \mathcal{H}_p.$$

The score operator has the form

$$V_n^\tau(\theta)(h) \equiv \frac{\partial}{\partial t} L_n(\theta_t)\Big|_{t=0} = V_{n,1}^\tau(\theta)(h_1) + V_{n,2}^\tau(\theta)(h_2),$$

where

$$V_{n,1}^{\tau}(\theta)(h_1) \equiv \mathbb{P}_n \left\{ h_1' Z N(\tau) - (1+\delta) \left[ \frac{h_1' Z e^{\beta' Z} A(U \wedge \tau)}{1 + e^{\beta' Z} A(U \wedge \tau)} \right] \right\},$$

and $V_{n,2}^{\tau}(\theta)(h_2)$ is defined by replacing $h_1$ with $h_2$ in (15.10). As mentioned earlier, we will need to utilize the dependence on $\tau$ at a later point. Now we have that the NPMLE $\hat{\theta}_n$ can be characterized as a zero of the map $\theta \mapsto \Psi_n(\theta) = V_n^{\tau}(\theta)$, and thus $\hat{\theta}_n$ is a Z-estimator with the estimating equation residing in $\ell^{\infty}(\mathcal{H}_{\infty})$.

The expectation of $\Psi_n$ is $\Psi \equiv PV^{\tau}$, where $V^{\tau}$ equals $V_1^{\tau}$ (i.e., $V_n^{\tau}$ with $n = 1$), with $X_1$ replaced by a generic observation $X$. Thus $V^{\tau}$ also satisfies $V_n^{\tau} = \mathbb{P}_n V^{\tau}$. The Gâteaux derivative of $\Psi$ at any $\theta_1 \in \Theta$ exists and is obtained by differentiating over the submodels $t \mapsto \theta_1 + t\theta$. This derivative is

$$\dot{\Psi}_{\theta_0}(h) \equiv \frac{\partial}{\partial t} \Psi(\theta_1 + t\theta) \Big|_{t=0} = -\theta(\sigma_{\theta_1}(h)),$$

where the operator $\sigma_{\theta} : \mathcal{H}_{\infty} \mapsto \mathcal{H}_{\infty}$ can be shown to be

(15.20)

$$\begin{aligned} \sigma_{\theta}(h) &= \begin{pmatrix} \sigma_{\theta}^{11} & \sigma_{\theta}^{12} \\ \sigma_{\theta}^{21} & \sigma_{\theta}^{22} \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} \\ &\equiv P \begin{pmatrix} \hat{\sigma}_{\theta}^{11} & \hat{\sigma}_{\theta}^{12} \\ \hat{\sigma}_{\theta}^{21} & \hat{\sigma}_{\theta}^{22} \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} \\ &\equiv P \hat{\sigma}_{\theta}(h), \end{aligned}$$

where

$$\begin{aligned} \hat{\sigma}_{\theta}^{11}(h_1) &= \hat{\xi}_{\theta} A(U) Z Z' h_1 \\ \hat{\sigma}_{\theta}^{12}(h_2) &= \hat{\xi}_{\theta} Z \int_0^{\tau} Y(s) h_2(s) dA(s) \\ \hat{\sigma}_{\theta}^{21}(h_1)(s) &= \hat{\xi}_{\theta} Y(s) Z' h_1 \\ \hat{\sigma}_{\theta}^{22}(h_2)(s) &= \frac{(1+\delta) e^{\beta' Z} Y(s) h_2(s)}{1 + e^{\beta' Z} A(U)} - \hat{\xi}_{\theta} e^{\beta' Z} \int_0^{\tau} Y(u) h_2(u) dA(u) Y(s), \end{aligned}$$

and

$$\hat{\xi}_{\theta} \equiv \frac{(1+\delta) e^{\beta' Z}}{\left[1 + e^{\beta' Z} A(U)\right]^2}.$$

We need to strengthen this Gâteaux differentiability of $\Psi$ to Fréchet differentiability, at least at $\theta = \theta_0$. This is accomplished in the following lemma:

LEMMA 15.8 *Under the given assumptions, the operator $\theta \mapsto \Psi(\theta)(\cdot)$, viewed as a map from $\ell^{\infty}(\mathcal{H}_p)$ to $\ell^{\infty}(\mathcal{H}_p)$, is Fréchet differentiable for each $p < \infty$ at $\theta = \theta_0$, with derivative $\theta \mapsto \dot{\Psi}_{\theta_0}(\theta) \equiv -\theta(\sigma_{\theta_0}(\cdot))$.*

Before giving the proof of this lemma, we note that we also need to verify that $\dot{\Psi}_{\theta_0}$ is continuously invertible. The following theorem establishes this plus a little more.

THEOREM 15.9 *Under the given conditions,* $\sigma_{\theta_0} : \mathcal{H}_\infty \mapsto \mathcal{H}_\infty$ *is continuously invertible and onto. Moreover,* $\dot{\Psi}_{\theta_0} : \overline{lin}\Theta \mapsto \overline{lin}\Theta$ *is also continuously invertible and onto, with inverse* $\theta \mapsto \dot{\Psi}_{\theta_0}^{-1}(\theta) \equiv -\theta(\sigma_{\theta_0}^{-1})$, *where* $\sigma_{\theta_0}^{-1}$ *is the inverse of* $\sigma_{\theta_0}$.

The "onto" property is not needed for the Z-estimator convergence theorem, but it will prove useful in Chapter 22 where we will revisit this example and show that $\hat{\theta}_n$ is asymptotically efficient. We conclude this section with the proofs of Lemma 15.8 and Theorem 15.9.

**Proof of Lemma 15.8.** By the smoothness of $\theta \mapsto \sigma_\theta(\cdot)$, we have

$$\lim_{t \downarrow 0} \sup_{\theta : \|\theta\|_{(p)} \leq 1} \sup_{h \in \mathcal{H}_p} \left| \int_0^\tau \theta \left( \sigma_{\theta_0 + ut\theta}(h) - \sigma_{\theta_0}(h) \right) du \right| = 0.$$

Thus

$$\sup_{h \in \mathcal{H}_p} \left| \Psi(\theta_0 + \theta)(h) - \Psi(\theta_0)(h) + \theta \left( \sigma_{\theta_0}(h) \right) \right| = o\left( \|\theta\|_{(p)} \right)$$

as $\|\theta\|_{(p)} \to 0$.□

**Proof of Theorem 15.9.** From the explicit form of the operator $\sigma_{\theta_0}$ defined above, we have that $\sigma_{\theta_0} = \sigma_1 + \sigma_2$, where

$$\sigma_1 \equiv \begin{pmatrix} I & 0 \\ 0 & g_0(\cdot) \end{pmatrix},$$

where $I$ is the $d \times d$ identity matrix,

$$g_0(s) = P\left[ \frac{(1+\delta) e^{\beta_0' Z} Y(s)}{1 + e^{\beta_0' Z} A_0(U)} \right],$$

and where $\sigma_2 \equiv \sigma_{\theta_0} - \sigma_1$. It is not hard to verify that $\sigma_2$ is a compact operator, i.e., that the range of $h \mapsto \sigma_2(h)$ over the unit ball in $\mathcal{H}_\infty$ lies within a compact set (see Exercise 15.6.6). Note that since $1/g_0$ has bounded total variation, we have for any $g = (g_1, g_2) \in \mathcal{H}$, that $g = \sigma_1(h)$, where $h = (g_1, g_2(\cdot)/g_0(\cdot)) \in \mathcal{H}$. Thus $\sigma_1$ is onto. It is also true that

$$\|g_0(\cdot) h_2(\cdot)\|_{\mathcal{H}} \geq \left( \inf_{s \in [0,\tau]} |g_0(s)| \right) \|h_2\|_{\mathcal{H}} \geq c_0 \|h_2\|_{\mathcal{H}},$$

and thus $\sigma_1$ is both continuously invertible and onto. If we can also verify that $\sigma_{\theta_0}$ is one-to-one, we then have by Lemma 6.17 that $\sigma_{\theta_0} = \sigma_1 + \sigma_2$ is both continuously invertible and onto.

We will now verify that $\sigma_{\theta_0}$ is one-to-one by showing that for any $h \in \mathcal{H}_\infty$, $\sigma_{\theta_0}(h) = 0$ implies that $h = 0$. Fix an $h \in \mathcal{H}_\infty$ for which $\sigma_{\theta_0}(h) = 0$, and define the one-dimensional submodel $t \mapsto \theta_{0t} = (\beta_{0t}, A_{0t}) \equiv (\beta_0, A_0) + t(h_1, \int_0^{(\cdot)} h_2(s)dA(s))$. Note that $\sigma_{\theta_0}(h) = 0$ implies

$$P\left\{ \frac{\partial^2}{(\partial t)^2} \ell_n(\theta_{0t}) \Big|_{t=0} \right\} = P[V^\tau(\theta_0)(h)]^2 = 0, \tag{15.21}$$

where we are using the original form of the likelihood $\ell_n$ instead of the modified form $L_n$ because the submodels $A_{0t}$ are differentiable for all $t$ small enough.

It can be verified that $V^u(\theta_0)(h)$ is a continuous time martingale over $u \in [0, \tau]$ (here is where we need the dependency of $V^\tau$ on $\tau$). The basic idea is that $V^u(\theta_0)(h)$ can be reexpressed as

$$\int_0^u \left( \frac{\dot{\lambda}_{\theta_0}(s)}{\lambda_{\theta_0}(s)} \right) (h_1' Z + h_2(s))\, dM(s),$$

where

$$\dot{\lambda}_{\theta_0} \equiv \frac{\partial}{\partial t} \lambda_{\theta_{0t}} \Big|_{t=0}, \quad \lambda_\theta(u) \equiv \frac{e^{\beta' Z} a_0(u)}{1 + e^{\beta' Z} A(u)},$$

and where $M(u) = N(u) - \int_0^u Y(s)\lambda_{\theta_0}(s)ds$ is a martingale since $\lambda_{\theta_0}$ is the correct hazard function for the failure time $T$ given $Z$. Thus $P[V^\tau(\theta_0)(h)]^2 = P[V^u(\tau)(\theta_0)(h)]^2 + P[V^\tau(\theta_0)(h) - V^u(\theta_0)(h)]^2$ for all $u \in [0, \tau]$, and hence $P[V^u(\theta_0)(h)]^2 = 0$ for all $u \in [0, \tau]$. Thus $V^u(\theta_0)(h) = 0$ almost surely, for all $u \in [0, \tau]$.

Hence, if we assume that the failure time $T$ is censored at some $U \in (0, \tau]$, we have almost surely that

$$\frac{e^{\beta_0' Z} \int_0^u (h_1' Z + h_2(s)) Y(s) dA_0(s)}{1 + \int_0^u e^{\beta_0' Z} Y(s) dA_0(s)} = 0,$$

for all $u \in [0, \tau]$. Hence $\int_0^u (h_1' Z + h_2(s)) Y(s) dA_0(s) = 0$ almost surely for all $u \in [0, \tau]$. Taking the derivative with respect to $u$ yields that $h_1' Z + h_2(u) = 0$ almost surely for all $u \in [0, \tau]$. This of course forces $h = 0$ since $\mathrm{var}[Z]$ is positive definite. Thus $\sigma_{\theta_0}$ is one-to-one since $h$ was an arbitrary choice satisfying $\sigma_{\theta_0}(h) = 0$. Hence $\sigma_{\theta_0}$ is continuously invertible and onto, and the first result of the theorem is proved.

We now prove the second result of the theorem. Note that since $\sigma_{\theta_0} : \mathcal{H}_\infty \mapsto \mathcal{H}_\infty$ is continuously invertible and onto, for each $p > 0$, there is a $q > 0$ such that $\sigma_{\theta_0}^{-1}(\mathcal{H}_q) \subset \mathcal{H}_p$. Fix $p > 0$, and note that

$$\begin{aligned} \inf_{\theta \in \mathrm{lin}\,\Theta} \frac{\|\theta(\sigma_{\sigma_0}(\cdot))\|_{(p)}}{\|\theta(\cdot)\|_{(p)}} &\geq \inf_{\theta \in \mathrm{lin}\,\Theta} \left[ \frac{\sup_{h \in \sigma_{\theta_0}^{-1}(\mathcal{H}_{(q)})} |\theta(\sigma_{\theta_0}^{-1}(h))|}{\|\theta\|_{(p)}} \right] \\ &= \inf_{\theta \in \Theta} \frac{\|\theta\|_{(q)}}{\|\theta\|_{(p)}} \geq \frac{q}{2p}. \end{aligned}$$

(See Exercise 15.6.4 to verify the last inequality.) Thus $\|\theta(\sigma_{\theta_0})\|_{(p)} \geq c_p\|\theta\|_{(p)}$, for all $\theta \in \lin\Theta$, where $c_p > 0$ depends only on $p$. Lemma 6.16, Part (i), now implies that $\theta \mapsto \theta(\sigma_{\theta_0})$ is continuously invertible. For any $\theta_1 \in \overline{\lin}\,\Theta$, we have $\theta_2(\sigma_{\theta_0}) = \theta_1$, where $\theta_2 = \theta_1(\sigma_{\theta_0}^{-1}) \in \overline{\lin}\,\Theta$. Thus $\theta \mapsto \theta(\sigma_{\theta_0})$ is also onto. Hence $\theta \mapsto \dot{\Psi}_{\theta_0}(\theta) = -\theta(\sigma_{\theta_0})$ is both continuously invertible and onto, and the theorem is proved.□

### *15.3.5 Weak Convergence and Bootstrap Validity*

Our approach to establishing weak convergence will be through verifying the conditions of Theorem 2.11 via the Donsker class result of Lemma 13.3. After establishing weak convergence, we will use a similar technical approach, but with some important differences, to obtain validity of a simple weighted bootstrap procedure.

Recall that $\Psi_n(\theta)(h) = \mathbb{P}_n V^\tau(\theta)(h)$, and note that $V^\tau(\theta)(h)$ can be expressed as

$$V^\tau(\theta)(h) = \int_0^\tau (h_1'Z + h_2(s))dN(s) - \int_0^\tau (h_1'Z + h_2(s))W(s;\theta)dA(s).$$

We now show that for any $0 < \epsilon < \infty$, $\mathcal{G}_\epsilon \equiv \{V^\tau(\theta)(h) : \theta \in \Theta_\epsilon, h \in \mathcal{H}_1\}$, where $\Theta_\epsilon \equiv \{\theta \in \Theta : \|\theta - \theta_0\|_{(1)} \leq \epsilon\}$ is $P$-Donsker. First, Lemma 15.7 tells us that $\{W(t;\theta) : t \in [0,\tau], \theta \in \Theta\}$ is Donsker. Second, it is easily seen that $\{h_1'Z + h_2(t) : t \in [0,\tau], h \in \mathcal{H}_1\}$ is also Donsker. Since the product of bounded Donsker classes is also Donsker, we have that $\{f_{t,\theta}(h) \equiv (h_1'Z + h_2(t))W(t;\theta) : t \in [0,\tau], \theta \in \Theta_\epsilon, h \in \mathcal{H}_1\}$ is Donsker. Third, consider the map $\phi : \ell^\infty([0,\tau] \times \Theta_\epsilon \times \mathcal{H}_1) \mapsto \ell^\infty(\Theta_\epsilon \times \mathcal{H}_1 \times \mathcal{A}_\epsilon)$ defined by

$$\phi(f_{\cdot,\theta}(h)) \equiv \int_0^\tau f_{s,\theta}(h)d\tilde{A}(s),$$

for $\tilde{A}$ ranging over $\mathcal{A}_\epsilon \equiv \{A \in \mathcal{A} : \sup_{t\in[0,\tau]} |A(t) - A_0(t)| \leq \epsilon\}$. Note that for any $\theta_1, \theta_2 \in \Theta_\epsilon$ and $h, \tilde{h} \in \mathcal{H}_1$,

$$\left|\phi(f_{\cdot,\theta_1}(h)) - \phi(f_{\cdot,\theta_2}(\tilde{h}))\right| \leq \sup_{t\in[0,\tau]} \left|f_{t,\theta_1}(h) - f_{t,\theta_2}(\tilde{h})\right| \times (A_0(\tau) + \epsilon).$$

Thus $\phi$ is continuous and linear, and hence $\{\phi(f_{\cdot,\theta}(h)) : \theta \in \Theta_\epsilon, h \in \mathcal{H}_1\}$ is Donsker by Lemma 15.10 below. Thus also

$$\left\{\int_0^\tau (h_1'Z + h_2(s))W(s;\theta)dA(s) : \theta \in \Theta_\epsilon, h \in \mathcal{H}_1\right\}$$

is Donsker. Since it not hard to verify that

$$\left\{\int_0^\tau (h_1'Z + h_2(s))dN(s) : h \in \mathcal{H}_1\right\}$$

is also Donsker, we now have that $\mathcal{G}_\epsilon$ is indeed Donsker as desired.

We now present the needed lemma and its proof before continuing:

LEMMA 15.10 *Suppose $\mathcal{F}$ is Donsker and $\phi : \ell^\infty(\mathcal{F}) \mapsto \mathbb{D}$ is continuous and linear. Then $\phi(\mathcal{F})$ is Donsker.*

**Proof.** Observe that $\mathbb{G}_n\phi(\mathcal{F}) = \phi(\mathbb{G}_n\mathcal{F}) \rightsquigarrow \phi(\mathbb{G}\mathcal{F}) = \mathbb{G}(\phi(\mathcal{F}))$, where the first equality follows from linearity, the weak convergence follows from the continuous mapping theorem, the second equality follows from a reapplication of linearity, and the meaning of the "abuse in notation" is obvious.□

We now have that both $\{V^\tau(\theta)(h) - V^\tau(\theta_0)(h) : \theta \in \Theta_\epsilon, h \in \mathcal{H}_1\}$ and $\{V^\tau(\theta_0)(h) : h \in \mathcal{H}_1\}$ are also Donsker. Thus $\sqrt{n}(\Psi_n(\theta_0) - \Psi(\theta_0)) \rightsquigarrow \mathbb{G}V^\tau(\theta_0)$ in $\ell^\infty(\mathcal{H}_1)$. Moreover, since it is not hard to show (see Exercise 15.6.7) that

$$\sup_{h\in\mathcal{H}_1} P\left(V^\tau(\theta)(h) - V^\tau(\theta_0)(h)\right)^2 \to 0, \qquad \text{as } \theta \to \theta_0, \tag{15.22}$$

Lemma 13.3 yields that

$$\left\|\sqrt{n}(\Psi_n(\theta) - \Psi(\theta)) - \sqrt{n}(\Psi_n(\theta_0) - \Psi(\theta_0))\right\|_{(1)} = o_P(1), \text{ as } \theta \to \theta_0.$$

Combining these results with Theorem 15.9, we have that all of the conditions of Theorem 2.11 are satisfied, and thus

$$\left\|\sqrt{n}\dot{\Psi}_{\theta_0}(\hat{\theta}_n - \theta_0) + \sqrt{n}(\Psi_n(\theta_0) - \Psi(\theta_0))\right\|_{(1)} = o_P(1) \tag{15.23}$$

and $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \mathcal{Z}_0 \equiv -\dot{\Psi}_{\theta_0}^{-1}(\mathbb{G}V^\tau(\theta_0))$ in $\ell^\infty(\mathcal{H}_1)$. We can observe from this result that $\mathcal{Z}_0$ is a tight, mean zero Gaussian process with covariance $P[\mathcal{Z}_0(h)\mathcal{Z}_0(\tilde{h})] = P\left[V^\tau(\theta_0)(\sigma_{\theta_0}^{-1}(h))V^\tau(\theta_0)(\sigma_{\theta_0}^{-1}(\tilde{h})\right]$, for any $h, \tilde{h} \in \mathcal{H}_1$. As pointed out earlier, this is in fact uniform convergence since any component of $\theta$ can be extracted via $\theta(h)$ for some $h \in \mathcal{H}_1$.

Now we will establish validity of a weighted bootstrap procedure for inference. Let $w_1, \ldots, w_n$ be positive, i.i.d., and independent of the data $X_1, \ldots, X_n$, with $0 < \mu \equiv Pw_1 < \infty$, $0 < \sigma^2 \equiv \text{var}(w_1) < \infty$, and $\|w_1\|_{2,1} < \infty$. Define the weighted bootstrapped empirical process $\tilde{\mathbb{P}}_n \equiv n^{-1}\sum_{i=1}^n (w_i/\bar{w})\Delta_{X_i}$, where $\bar{w} \equiv n^{-1}\sum_{i=1}^n w_i$ and $\Delta_{X_i}$ is the empirical measure for the observation $X_i$. This particular bootstrap was introduce in Section 2.2.3. Let $\tilde{L}_n(\theta)$ be $L_n(\theta)$ but with $\mathbb{P}_n$ replaced by $\tilde{\mathbb{P}}_n$, and let $\tilde{\Psi}_n$ be $\Psi_n$ but with $\mathbb{P}_n$ replaced by $\tilde{\mathbb{P}}_n$. Define $\tilde{\theta}_n$ to be the maximizer of $\theta \mapsto \tilde{L}_n(\theta)$. The idea is, after conditioning on the data sample $X_1, \ldots, X_n$, to compute $\tilde{\theta}_n$ for many replications of the weights $w_1, \ldots, w_n$ to form confidence intervals for $\theta_0$. We want to show that

$$\sqrt{n}(\mu/\sigma)(\tilde{\theta}_n - \hat{\theta}_n) \underset{w}{\overset{\text{P}}{\rightsquigarrow}} \mathcal{Z}_0. \tag{15.24}$$

We first study the unconditional properties of $\tilde{\theta}_n$. Note that for maximizing $\theta \mapsto \tilde{L}_n(\theta)$ and for zeroing $\theta \mapsto \tilde{\Psi}_n$, we can temporarily drop the

$\bar{w}$ factor since neither the maximizer nor zero of a function is modified when multiplied by a positive constant. Let $w$ be a generic version of $w_1$, and note that if a class of functions $\mathcal{F}$ is Glivenko-Cantelli, then so also is the class of functions $w \cdot \mathcal{F}$ via Theorem 10.13. Likewise, if the class $\mathcal{F}$ is Donsker, then so is $w \cdot \mathcal{F}$ via the multiplier central limit theorem, Theorem 10.1. Also, $Pwf = \mu Pf$, trivially. What this means, is that the arguments in Sections 15.3.2 and 15.3.3 can all be replicated for $\tilde{\theta}_n$ with only trivial modifications. This means that $\tilde{\theta}_n \stackrel{\text{as*}}{\to} \theta_0$.

Now, reinstate the $\bar{w}$ everywhere, and note by Corollary 10.3, we can verify that both $\sqrt{n}(\tilde{\Psi} - \Psi)(\theta_0) \rightsquigarrow (\sigma/\mu)\mathbb{G}_1 V^\tau(\theta_0) + \mathbb{G}_2 V^\tau(\theta_0)$, where $\mathbb{G}_1$ and $\mathbb{G}_2$ are independent Brownian bridge random measures, and

$$\left\| \sqrt{n}(\tilde{\Psi}_n(\tilde{\theta}_n) - \Psi(\tilde{\theta}_n)) - \sqrt{n}(\tilde{\Psi}_n(\theta_0) - \Psi(\theta_0)) \right\|_{(1)} = o_P(1).$$

Thus reapplication of Theorem 2.11 yields that

$$\left\| \sqrt{n}\dot{\Psi}_{\theta_0}(\tilde{\theta}_n - \theta_0) + \sqrt{n}(\tilde{\Psi}_n - \Psi)(\theta_0) \right\|_{(1)} = o_P(1).$$

Combining this with (15.23), we obtain

$$\left\| \sqrt{n}\dot{\Psi}_{\theta_0}(\tilde{\theta}_n - \hat{\theta}_n) + \sqrt{n}(\tilde{\Psi}_n - \Psi_n)(\theta_0) \right\|_{(1)} = o_P(1).$$

Now, using the linearity of $\dot{\Psi}_{\theta_0}$, the continuity of $\dot{\Psi}_{\theta_0}^{-1}$, and the bootstrap central limit theorem, Theorem 2.6, we have the desired result that $\sqrt{n}(\mu/\sigma)(\tilde{\theta}_n - \hat{\theta}_n) \underset{w}{\overset{\text{P}}{\rightsquigarrow}} \mathcal{Z}_0$. Thus the proposed weighted bootstrap is valid.

We also note that it is not clear how to verify the validity of the usual nonparametric bootstrap, although its validity probably does hold. The key to the relative simplicity of the theory for the proposed weighted bootstrap is that Glivenko-Cantelli and Donsker properties of function classes are not altered after multiplying by independent random weights satisfying the given moment conditions. We also note that the weighted bootstrap is computationally simple, and thus it is quite practical to generate a reasonably large number of replications of $\tilde{\theta}_n$ to form confidence intervals. This is demonstrated numerically in Kosorok, Lee and Fine (2004).

## 15.4 Testing for a Change-point

Recall the change-point model example of Section 14.5.1 and consider testing the null hypothesis $H_0 : \alpha = \beta$. Under this null, the change-point parameter $\zeta$ is not identifiable, and thus $\hat{\zeta}_n$ is not consistent. This means that testing $H_0$ is an important concern, since it is unlikely we would know in advance whether $H_0$ were true. The statistic we propose using is $T_n \equiv \sup_{\zeta \in [a,b]} |U_n(\zeta)|$, where

$$U_n(\zeta) \equiv \frac{\sqrt{n\hat{F}_n(\zeta)(1-\hat{F}_n(\zeta))}}{\hat{\sigma}_n}\left(\frac{\sum_{i=1}^n 1\{Z_i \le \zeta\}Y_i}{n\hat{F}_n(\zeta)} - \frac{\sum_{i=1}^n 1\{Z_i > \zeta\}Y_i}{n(1-\hat{F}_n(\zeta))}\right),$$

$\hat{\sigma}_n^2 \equiv n^{-1}\sum_{i=1}^n \left(Y_i - \hat{\alpha}_n 1\{Z_i \le \hat{\zeta}_n\} - \hat{\beta}_n 1\{Z_i > \hat{\zeta}_n\}\right)^2$, and where $\hat{F}_n(t) \equiv \mathbb{P}_n 1\{Z \le t\}$. We will study the asymptotic limiting behavior of this statistic under the sequence of contiguous alternative hypotheses $H_{1n} : \beta = \alpha + \eta/\sqrt{n}$, where the distribution of $Z$ and $\epsilon$ does not change with $n$. Thus $Y_i = \epsilon_i + \alpha_0 + (\eta/\sqrt{n})1\{Z_i > \zeta_0\}$, $i = 1, \ldots, n$.

We will first show that under $H_{1n}$,

$$U_n(\zeta) = a_0(\zeta)B_n(\zeta) + \nu_0(\zeta) + r_n(\zeta), \tag{15.25}$$

where $a_0(\zeta) \equiv \sigma^{-1}\sqrt{F(\zeta)(1-F(\zeta))}$, $F(\zeta) \equiv P1\{Z \le \zeta\}$,

$$\begin{aligned} B_n(\zeta) &\equiv \frac{\sqrt{n}\mathbb{P}_n[1\{\zeta_0 < Z \le \zeta\}\epsilon]}{F(\zeta)} - \frac{\sqrt{n}\mathbb{P}_n[1\{Z > \zeta \vee \zeta_0\}\epsilon]}{1-F(\zeta)}, \\ \nu_0(\zeta) &\equiv \eta a_0(\zeta)\left(\frac{P[\zeta_0 < Z \le \zeta]}{F(\zeta)} - \frac{P[Z > \zeta \vee \zeta_0]}{1-F(\zeta)}\right), \end{aligned}$$

and $\sup_{\zeta\in[a,b]}|r_n(\zeta)| \xrightarrow{\mathrm{P}} 0$. The first step is to note that $\hat{F}_n$ is uniformly consistent on $[a,b]$ for $F$ and that both $\inf_{\zeta\in[a,b]}F(\zeta) > 0$ and $\inf_{\zeta\in[a,b]}(1-F(\zeta)) > 0$. Since $\eta/\sqrt{n} \to 0$, we also have that both

$$V_n(\zeta) \equiv \frac{\sum_{i=1}^n 1\{Z_i \le \zeta\}Y_i}{\sum_{i=1}^n 1\{Z_i \le \zeta\}}, \text{ and } W_n(\zeta) \equiv \frac{\sum_{i=1}^n 1\{Z_i \ge \zeta\}Y_i}{\sum_{i=1}^n 1\{Z_i \ge \zeta\}}$$

are uniformly consistent, over $\zeta \in [a,b]$, for $\alpha_0$ by straightforward Glivenko-Cantelli arguments. Since $\hat{\zeta}_n \in [a,b]$ with probability 1 by assumption, we have $\hat{\sigma}_n^2 \xrightarrow{\mathrm{P}} \sigma^2$ even though $\hat{\zeta}_n$ may not be consistent. Now the fact that both $\mathcal{F}_1 \equiv \{1\{Z \le \zeta\} : \zeta \in [a,b]\}$ and $\mathcal{F}_2 \equiv \{1\{Z > \zeta\} : \zeta \in [a,b]\}$ are Donsker yields the final conclusion, after some simple calculations.

Reapplying the fact that $\mathcal{F}_1$ and $\mathcal{F}_2$ are Donsker yields that $a_0(\zeta)B_n(\zeta) \rightsquigarrow \mathcal{B}_0(\zeta)$ in $\ell^\infty([a,b])$, where $\mathcal{B}_0$ is a tight, mean zero Gaussian process with covariance

$$H_0(\zeta_1,\zeta_2) \equiv \sqrt{\frac{F(\zeta_1 \wedge \zeta_2)(1-F(\zeta_1 \vee \zeta_2))}{F(\zeta_1 \vee \zeta_2)(1-F(\zeta_1 \wedge \zeta_2))}}.$$

Thus, from (15.25), we have the $U_n \rightsquigarrow \mathcal{B}_0 + \nu_0$, it is easy to verify that $\nu_0(\zeta_0) \neq 0$. Hence, if we use critical values based on $\mathcal{B}_0$, the statistic $T_n$ will asymptotically have the correct size under $H_0$ as well as have arbitrarily large power under $H_{1n}$ as $|\eta|$ gets larger.

Thus what we need now is a computationally easy method for obtaining the critical values of $\sup_{\zeta\in[a,b]}|\mathcal{B}_0(\zeta)|$. Define

$$A_i^n(\zeta) \equiv \sqrt{\hat{F}_n(\zeta)(1-\hat{F}_n(\zeta))}\left(\frac{1\{Z_i \le \zeta\}}{\hat{F}_n(\zeta)} - \frac{1\{Z_i > \zeta\}}{1-\hat{F}_n(\zeta))}\right),$$

and consider the weighted "bootstrap" $\mathcal{B}_n(\zeta) \equiv n^{-1/2}\sum_{i=1}^n A_i^n(\zeta)w_i$, where $w_1, \ldots, w_n$ are i.i.d. standard normals independent of the data. The continuous mapping theorem applied to the following lemma verifies that this bootstrap will satisfy $\sup_{\zeta\in[a,b]}|\mathcal{B}_n(\zeta)| \underset{w}{\overset{\mathrm{P}}{\rightsquigarrow}} \sup_{\zeta\in[a,b]}|\mathcal{B}_0(\zeta)|$ and thus can be used to obtain the needed critical values:

LEMMA 15.11 *$\mathcal{B}_n \underset{w}{\overset{\mathrm{P}}{\rightsquigarrow}} \mathcal{B}_0$ in $\ell^\infty([a,b])$.*

Before ending this section with the proof of this lemma, we note that the problem of testing a null hypothesis under which some of the parameters are no longer identifiable is quite challenging (see Andrews, 2001), especially when infinite-dimensional parameters are present. An example of the latter setting is in transformation regression models with a change-point (Kosorok and Song, 2007). Obtaining the critical value of the appropriate statistic for these models is much more difficult than it is for the simple example we have presented in this section.

**Proof of Lemma 15.11.** Let $\mathcal{G}$ be the class of nondecreasing functions $G : [a,b] \mapsto [a_0, b_0]$ of $\zeta$, where $a_0 \equiv F(a)/2$ and $b_0 \equiv 1/2 + F(b)/2$, and note that by the uniform consistency of $\hat{F}_n$, $\hat{F}_n \in \mathcal{G}$ with probability arbitrarily close to 1 for all $n$ large enough. Also note that

$$\left\{\sqrt{G(\zeta)(1-G(\zeta))}\left(\frac{1\{Z \le \zeta\}}{G(\zeta)} - \frac{1\{Z > \zeta\}}{1-G(\zeta)}\right) : \zeta \in [a,b], G \in \mathcal{G}\right\}$$

is Donsker, since $\{a\mathcal{F} : a \in K\}$ is Donsker for any compact $K \subset \mathbb{R}$ and any Donsker class $\mathcal{F}$. Thus by the multiplier central limit theorem, Theorem 10.1, we have that $\mathcal{B}_n$ converges weakly to a tight Gaussian process unconditionally. Thus we know that $\mathcal{B}_n$ is asymptotically tight conditionally, i.e., we know that

$$\mathrm{P}^*\left(\sup_{\zeta_1,\zeta_2\in[a,b]:|\zeta_1-\zeta_2|\le\delta_n} |\mathcal{B}_n(\zeta_1) - \mathcal{B}_n(\zeta_2)| > \tau \,\middle|\, X_1, \ldots, X_n\right) = o_P(1),$$

for all $\tau > 0$ and all sequences $\delta_n \downarrow 0$. All we need to verify now is that the finite dimensional distributions of $\mathcal{B}_n$ converge to the appropriate limiting multivariate normal distributions conditionally.

Since $w_1, \ldots, w_n$ are i.i.d. standard normal, it is easy to see that $\mathcal{B}_n$ is conditionally a Gaussian process with mean zero and, after some algebra (see Exercise 15.6.8), with covariance

$$\hat{H}_n(\zeta_1,\zeta_2) \equiv \sqrt{\frac{\hat{F}_n(\zeta_1\wedge\zeta_2)(1-\hat{F}_n(\zeta_1\vee\zeta_2))}{\hat{F}_n(\zeta_1\vee\zeta_2)(1-\hat{F}_n(\zeta_1\wedge\zeta_2))}}. \tag{15.26}$$

Since $\hat{H}_n$ can easily be shown to converge to $H_0(\zeta_1,\zeta_2)$ uniformly over $[a,b]\times[a,b]$, the proof is complete.□

## 15.5 Large $p$ Small $n$ Asymptotics for Microarrays

In this section, we present results for the "large p, small n" paradigm (West, 2003) that arises in microarray studies, image analysis, high throughput molecular screening, astronomy, and in many other high dimensional applications. To solidify our discussion, we will focus on an idealized version of microarray data. The material for this section comes primarily from Kosorok and Ma (2007), which can be referred to for additional details and which contains a brief review of statistical methods for microarrays. Microarrays are capable of monitoring the gene expression of thousands of genes and have become both important and routine in biomedical research. The basic feature of a simple microarray experiment is that data from a small number $n$ of i.i.d. microarrays are collected, with each microarray having measurements on a large number $p$ of genes. Typically, $n$ is around 20 or less (usually, in fact, much less) whereas $p$ is easily in the thousands. The data is usually aggregated across the $n$ arrays to form test statistics for each of the $p$ genes, resulting in large scale multiple testing. False discovery rate (FDR) methods (see Benjamini and Hochberg, 1995) are used to account for this multiplicity in order to successfully identify which among thousands of monitored genes are significantly differentially expressed.

For FDR methods to be valid for identifying differentially expressed genes, the p-values for the non-differentially expressed genes must simultaneously have uniform distributions marginally. While this is feasible for permutation based p-values, it is unclear whether such uniformity holds for p-values based on asymptotic approximations. For instance, suppose that we wish to use t-tests for each of the $p$ genes and to compute approximate p-values based on the normal approximation for simplicity. If the data is not normally distributed, we would have to rely on the central limit theorem. Unfortunately, it is unclear whether this will work for all of the tests simultaneously. The issue is that the number of p-values involved goes to infinity and intuition suggests that at least some of the p-values should behave erratically. In this section, we will show the somewhat surprising result that, under arbitrary dependency structures across genes, the p-values are in fact simultaneously valid when using the normal approximation for t-tests. The result is essentially a very high dimensional central limit theorem.

To further clarify ideas, consider a simple one-sample cDNA microarray study. Note that this data setting and the results that follow can be easily extended to incorporate more complex designs. Studies using Affymetrix genechip data can also be included in the same framework after some modification. Denote $Y_{ij}$ as the background-corrected log-ratios, for array $i = 1, \ldots, n$ and gene $j = 1, \ldots, p$. Consider the following simple linear model:

$$Y_{ij} = \mu_j + \epsilon_{ij}, \tag{15.27}$$

where $\mu_j$ are the fixed gene effects and $\epsilon_{ij}$ are mean zero random errors. For simplicity of exposition, we have omitted other potentially important terms in our model, such as array-specific intensity normalization and possible print-tip effects. We note, however, that the theory we present in this paper can extend readily to these richer models, and, in fact has been extended to address normalization in the original Kosorok and Ma (2007) paper.

Under the simple model (15.27), permutation tests can be used. However, when normalization is needed and the residual errors are not Gaussian, it may be convenient to use the normal approximation to compute the p-values for the t-tests associated with the sample mean $\bar{X}_j$ for each gene $j = 1, \ldots, p$. In this section, we wish to study inference for these sample means when the number of arrays $n \to \infty$ slowly while the number of genes $p >> n$. This is essentially the asymptotic framework considered in van der Laan and Bryan (2001) who show that provided the range of expression levels is bounded, the sample means consistently estimate the mean gene effects uniformly across genes whenever $\log p = o(n)$. We extend the results of van der Laan and Bryan (2001) in two important ways. First, uniform consistency results are extended to general empirical distribution functions. This consistency will apply to sample means and will also apply to other functionals of the empirical distribution function such as sample medians (this is explored more in the Kosorok and Ma, 2007, paper). Second, a precise Brownian bridge approximation to the empirical distribution function is developed and utilized to establish uniform validity of marginal p-values based on the normal approximation. More specifically, we develop a central limit theorem for the large $p$ small $n$ setting which holds provided $\log p = o(n^{1/2})$.

An important consequence of these results is that approximate p-values based on normalized gene expression data can be validly applied to FDR methods for identifying differentially expressed genes. We refer to this kind of asymptotic regime as "marginal asymptotics" (see also Kosorok and Ma, 2005) because the focus of the inference is at the marginal (gene) level, even though the results are uniformly valid over all genes. Qualitatively, the requirement that $\log p = o(n^{1/2})$ seems to be approximately the correct order of asymptotics for microarray experiments with a moderate number, say $\sim$ 20–30, of replications. The main technical tools we use from empirical processes include maximal inequalities (Section 8.1), and a specialized Hungarian construction for the empirical distribution function.

We first discuss what is required of p-value approximations in order for the FDR procedure to be valid asymptotically. We then present the main theoretical results for the empirical processes involved, and then close this section with a specialization of these results to estimation and inference based on sample means.

### *15.5.1 Assessing P-Value Approximations*

Suppose we have $p$ hypothesis tests with p-values $q_{(p)} \equiv \{q_1, \ldots, q_p\}$ but only know the estimated p-values $\hat{q}_{(p)} \equiv \{\hat{q}_1, \ldots, \hat{q}_p\}$. An important question is how accurate must $\hat{q}_{(p)}$ be in order for inference based on $\hat{q}_{(p)}$ to be asymptotically equivalent to inference based on $q_{(p)}$? Specifically, the chief hypothesis testing issue is controlling the FDR asymptotically in $p$. To fix ideas, suppose the indices $J_p = \{1, \ldots, p\}$ for the hypothesis tests are divided into two groups, $J_{0p}$ and $J_{1p}$, where some null hypotheses hold for all $j \in J_{0p}$ and some alternatives hold for all $j \in J_{1p}$. We will assume that $q_j$ is uniformly distributed for all $j \in J_{0p}$ and that $q_j$ has distribution $F_1$ for all $j \in J_{1p}$, where $F_1(t) \geq t$ for all $t \in [0,1]$ and $F_1$ is strictly concave with $\lim_{t \downarrow 0} F_1(t)/t = \infty$. Let $\lambda_p \equiv \#J_{0p}/p$ be the proportion of true null hypotheses, and assume $\lambda_p \to \lambda_0 \in (0,1]$, as $p \to \infty$. Also let $\tilde{F}_p(t) \equiv p^{-1} \sum_{j=1}^{p} 1\{q_j \leq t\}$, where $1\{A\}$ is the indicator of $A$, and assume $\tilde{F}_p(t)$ converges uniformly in $t$ to $F_0(t) \equiv \lambda_0 t + (1-\lambda_0)F_1(t)$.

The estimate of FDR proposed by Storey (2002) (see also Genovese and Wasserman, 2002) for a p-value threshold of $t \in [0,1]$ is $\widetilde{FDR}_l(t) \equiv \tilde{\lambda}(l)t/(\tilde{F}_p(t) \vee (1/p))$, where $\tilde{\lambda}(l) \equiv (1 - \tilde{F}_p(l))/(1-l)$ is a conservative estimate of $\lambda_0$, in that $\tilde{\lambda}(l) \to \lambda_*$ in probability, where $\lambda_0 \leq \lambda_* \leq 1$, and where $a \vee b$ denotes the maximum of $a$ and $b$. The quantity $l$ is the tuning parameter and is constrained to be in $(0,1)$ with decreasing bias as $l$ gets closer to zero. Because of the upward bias in $\tilde{\lambda}(l)$, if $\tilde{\lambda}(l)$ is distinctly $< 1$, then one can be fairly confident that $\lambda_0 < 1$.

Assume $\lambda_0 < 1$. The asymptotic FDR for the procedure rejecting all hypotheses corresponding to indices with $p_j \leq t$ is $r_0(t) \equiv \lambda_0 t/(\lambda_0 t + (1-\lambda_0)F_1(t))$. Storey, Taylor and Siegmund (2004) demonstrate that under fairly general dependencies among the p-values $q_{(p)}$, $\tilde{F}_p(t)$ converges to $F_0(t)$, and thus $\widetilde{FDR}_l(t)$ converges in probability to $r_*(t) \equiv (\lambda_*/\lambda_0)r_0(t)$. Our assumptions on $F_1$ ensure that $r_0(t)$ is monotone increasing with derivative $\dot{r}_0(t)$ bounded by $(4\delta)^{-1}$. Thus, for each $\rho \in [0, \lambda_*]$, there exists a $t \in [0,1]$ with $r_*(t) = \rho$ and $r_0(t) \leq \rho$. Thus using $\widetilde{FDR}_l(t)$ to control FDR is asymptotically valid, albeit conservative.

Suppose all we have available is $\hat{q}_{(p)}$. Now we estimate $F_0$ with $\hat{F}_p(t) = p^{-1} \sum_{j=1}^{p} 1\{\hat{q}_j \leq t\}$ and $\lambda_*$ with $\hat{\lambda}_l(t) \equiv (1 - \hat{F}_p(l))/(1-l)$. The previous results will all hold for $\widehat{FDR}_l(t) \equiv \hat{\lambda}(l)t/(\hat{F}_p(t) \vee (1/p))$, provided $\hat{F}_p$ is uniformly consistent for $F_0$. We now show that a sufficient condition for this is $\max_{1 \leq j \leq p} |\hat{q}_j - q_j| \to 0$ in probability. Under this condition, there exists a positive sequence $\epsilon_p \downarrow 0$ such that $P\left(\max_{1 \leq j \leq p} |\hat{q}_{(p)} - q_{(p)}| > \epsilon_p\right) \to 0$ in probability. Accordingly, we have with probability tending to one that for any $t \in [\epsilon_p, 1-\epsilon_p]$, $\tilde{F}_p(t - \epsilon_p) \leq \hat{F}_p(t) \leq \tilde{F}_p(t + \epsilon_p)$. Thus, by continuity of $F_0$, uniform consistency of $\hat{F}_p$ follows from uniform consistency of $\tilde{F}_p$.

In summary, the above procedure for controlling FDR is asymptotically valid when $\lambda_0 < 1$, provided $\max_{1 \le j \le p} |\hat{q}_{(p)} - q_{(p)}|$ goes to zero in probability. However, This result does not hold when $\lambda_0 = 1$. This is discussed in greater detail in Kosorok and Ma (2007) who also present methods of addressing this issue which we do not pursue further here.

### *15.5.2 Consistency of Marginal Empirical Distribution Functions*

For each $n \ge 1$, let $X_{1(n)}, \ldots, X_{n(n)}$ be a sample of i.i.d. vectors (eg., microarrays) of length $p_n$, where the dependence within vectors is allowed to be arbitrary. Denote the $j$th component (eg., gene) of the $i$th vector $X_{ij(n)}$, i.e., $X_{i(n)} = (X_{i1(n)}, \ldots, X_{ip_n(n)})'$. Also let the marginal distribution of $X_{1j(n)}$ be denoted $F_{j(n)}$, and let $\hat{F}_{j(n)}(t) = n^{-1} \sum_{i=1}^{n} 1\{X_{ij(n)} \le t\}$, for all $t \in \mathbb{R}$ and each $j = 1, \ldots, p_n$.

The main results of this subsection are Theorems 15.12 and 15.13 below. The two theorems are somewhat surprising, high dimensional extensions of two classical univariate results for empirical distribution functions: the celebrated Dvoretzky, Kiefer and Wolfowitz (1956) inequality as refined by Massart (1990) and the celebrated Komlós, Major and Tusnády (1976) Hungarian (KMT) construction as refined by Bretagnolle and Massart (1989). The extensions utilize maximal inequalities based on Orlicz norms (see Section 8.1). The proofs are given at the end of this subsection.

The first theorem yields simultaneous consistency of all $\hat{F}_{j(n)}$s:

THEOREM 15.12 *For a universal constant $0 < c_0 < \infty$ and all $n, p_n \ge 2$,*

$$\left\| \max_{1 \le j \le p_n} \left\| \hat{F}_{j(n)} - F_{j(n)} \right\|_\infty \right\|_{\psi_2} \le c_0 \sqrt{\frac{\log p_n}{n}}. \tag{15.28}$$

*In particular, if $n \to \infty$ and $\log p_n = o(n)$, then the left side of (15.28) $\to 0$.*

Note that the rate on the right-side of (15.28) is sharp, in the sense that there exist sequences of data sets, where $(\log p_n / n)^{-1/2} \times \max_{1 \le j \le p_n} \|\hat{F}_{j(n)} - F_{j(n)}\|_\infty \to c > 0$, in probability, as $n \to \infty$. In particular, this is true if the genes are all independent, $n, p_n \to \infty$ with $\log p_n = o(n)$, and $c = 1/2$.

The second theorem shows that the standardized empirical processes $\sqrt{n}(\hat{F}_{j(n)} - F_{j(n)})$ can be simultaneously approximated by Brownian bridges in a manner which preserves the original dependency structure in the data. For example, if the original data has *weak dependence*, as defined in Storey, Taylor and Siegmund (2004), then so will the approximating Brownian bridges. An example of weak dependence is the $m$-dependence described in Section 11.6. Let $\mathcal{F}_{j(n)}$ be the smallest $\sigma$-field making all of $X_{1j(n)}, \ldots, X_{nj(n)}$ measurable, $1 \le j \le p_n$, and let $\mathcal{F}_n$ be the smallest $\sigma$-field making all of $\mathcal{F}_{1(n)}, \ldots, \mathcal{F}_{p_n(n)}$ measurable.

THEOREM 15.13 *For universal constants* $0 < c_1, c_2 < \infty$ *and all* $n, p_n \geq 2$,

(15.29)
$$\left\| \max_{1 \leq j \leq p_n} \left\| \sqrt{n}(\hat{F}_{j(n)} - F_{j(n)}) - B_{j(n)}(F_{j(n)}) \right\|_\infty \right\|_{\psi_1} \leq \frac{c_1 \log n + c_2 \log p_n}{\sqrt{n}},$$

*for some stochastic processes* $B_{1(n)}, \ldots, B_{p_n(n)}$ *which are conditionally independent given* $\mathcal{F}_n$ *and for which each* $B_{j(n)}$ *is a standard Brownian bridge with conditional distribution given* $\mathcal{F}_n$ *depending only on* $\mathcal{F}_{j(n)}$, $1 \leq j \leq p_n$.

Before giving the promised proofs, we note that while the above results are only used in the next subsection for developing inference based on the mean, these results extend nicely to median and other robust estimation methods (see Kosorok and Ma, 2007, and Ma, Kosorok, Huang, et al., 2006).

*Proof of Theorem 15.12.* Define $V_{j(n)} \equiv \sqrt{n}\|\hat{F}_{j(n)} - F_{j(n)}\|_\infty$, and note that by Corollary 1 of Massart (1990), $\mathrm{P}(V_{j(n)} > x) \leq 2e^{-2x^2}$, for all $x \geq 0$ and any distribution $F_{j(n)}$. This inequality is a refinement of the celebrated result of Dvoretzky, Kiefer and Wolfowitz (1956), given in their Lemma 2, and the extension to distributions with discontinuities is standard. Using Lemma 8.1, we obtain $\|V_{j(n)}\|_{\psi_2} \leq \sqrt{3/2}$ for all $1 \leq j \leq p_n$. Now, by Lemma 8.2 combined with the fact that

$$\limsup_{x,y \to \infty} \frac{\psi_2(x)\psi_2(y)}{\psi_2(xy)} = 0,$$

we have that there exists a universal constant $c_* < \infty$ with

$$\left\| \max_{1 \leq j \leq p_n} V_{j(n)} \right\|_{\psi_2} \leq c_* \sqrt{\log(1 + p_n)} \sqrt{3/2}$$

for all $n \geq 1$. The desired result now follows for the constant $c_0 = \sqrt{3}c_*$, since $\log(k+1) \leq 2 \log k$ for any $k \geq 2$.□

*Proof of Theorem 15.13.* Let $U_j$, $j = 1, \ldots, p_n$, be i.i.d. uniform random variables independent of the data. Then, by Theorem 15.14 below, we have for each $1 \leq j \leq p_n$ that there exists a measurable map $g_{j(n)} : \mathbb{R}^n \times [0,1] \mapsto C[0,1]$ where $B_{j(n)} = g_{j(n)}(X_{1j(n)}, \ldots, X_{nj(n)}, U_j)$ is a Brownian bridge with

(15.30)
$$P\left(\sqrt{n}\left\|\sqrt{n}(\hat{F}_{j(n)} - F_{j(n)}) - B_{j(n)}(F_{j(n)})\right\|_\infty > x + 12\log n\right) \leq 2e^{-x/6},$$

for all $x \geq 0$. Note that this construction generates an ensemble of Brownian bridges $B_{1(n)}, \ldots, B_{p_n(n)}$ that may be dependent when the components in $X_{1(n)} = (X_{11(n)}, \ldots, X_{1p_n(n)})'$ are dependent. However, each $B_{j(n)}$ only

depends on the information contained in $\mathcal{F}_{j(n)}$ and the independent uniform random variable $U_j$. Thus $B_{j(n)}$ depends on $\mathcal{F}_n$ only through the information contained in $\mathcal{F}_{j(n)}$, and the ensemble of Brownian bridges is conditionally independent given $\mathcal{F}_n$. Note also the validity of (15.30) for all $n \geq 2$.

Let $V_{j(n)} = \left( (\sqrt{n}/(\log n)) \left\| \sqrt{n}(\hat{F}_{j(n)} - F_{j(n)}) - B_{j(n)}(F_{j(n)}) \right\|_\infty - 12 \right)^+$, where $u^+$ is the positive part of $u$. By Lemma 8.1, Expression (15.30) implies that $\|V_{j(n)}\|_{\psi_1} \leq 18/\log n$. Reapplying the result that $\log(k+1) \leq 2\log k$ for any $k \geq 2$, we now have, by the fact that $\limsup_{x,y\to\infty} \psi_1(x)\psi_1(y)/\psi_1(xy) = 0$ combined with Lemma 8.2, that there exists a universal constant $0 < c_2 < \infty$ for which $\left\| \max_{1\leq j\leq p_n} V_{j(n)} \right\|_{\psi_1} \leq c_2 \log p_n/(\log n)$. Now (15.29) follows, for $c_1 = 12$, from the definition of $V_{j(n)}$.□

THEOREM 15.14 *For $n \geq 2$, let $Y_1, \ldots, Y_n$ be i.i.d. real random variables with distribution $G$ (not necessarily continuous), and let $U_0$ be a uniform random variable independent of $Y_1, \ldots, Y_n$. Then there exists a measurable map $g_n : \mathbb{R}^n \times [0,1] \mapsto C[0,1]$ such that $B = g_n(Y_1, \ldots, Y_n, U_0)$ is a standard Brownian bridge satisfying, for all $x \geq 0$,*

(15.31)
$$P\left( \sqrt{n} \left\| \sqrt{n}(\hat{G}_n - G) - B(G) \right\|_\infty > x + 12\log n \right) \quad \leq \quad 2e^{-x/6},$$

*where $\hat{G}_n$ is the empirical distribution of $Y_1, \ldots, Y_n$.*

*Proof.* By Theorem 20.4 of Billingsley (1995), there exists a measurable $h_0 : [0,1] \mapsto [0,1]^2$ such that $(U_1, U_2) \equiv h_0(U_0)$ is a pair of independent uniforms. Moreover, standard arguments yield the existence of a function $h_n : \mathbb{R}^n \times [0,1] \mapsto [0,1]^n$ such that $(V_1, \ldots, V_n) \equiv h_n(Y_1, \ldots, Y_n, U_1)$ is a sample of i.i.d. uniforms and $(Y_1, \ldots, Y_n) = (\psi(V_1), \ldots, \psi(V_n))$, where $\psi(u) \equiv \inf\{x : G(x) \geq u\}$ (see Exercise 15.6.9). $U_1$ is needed to handle possible discontinuities in $G$.

Let $\hat{H}_n$ be the empirical distribution for $V_1, \ldots, V_n$, and note that

(15.32)
$$\sqrt{n}(\hat{H}_n(G(x)) - G(x)) \quad = \quad \sqrt{n}(\hat{G}_n(x) - G(x)), \quad \forall x \in \mathbb{R}.$$

by design. Now by the Hungarian construction (Theorem 1) of Bretagnolle and Massart (1989), there exists a Brownian bridge $B$ depending only on $V_1, \ldots, V_n$ and $U_2$ such that

$$P\left( \sqrt{n} \sup_{u\in[0,1]} \left| \sqrt{n}(\hat{H}_n(u) - u) - B(u) \right| > x + 12\log n \right) \leq 2e^{-x/6},$$

for all $x \geq 0$, and thus by (15.32),

$$P\left(\sqrt{n}\left\|\sqrt{n}(\hat{G}_n - G) - B(G)\right\|_\infty > x + 12\log n\right) \leq 2e^{-x/6}, \ \forall x \geq 0. \tag{15.33}$$

By Lemma 15.15 below, we can take $B$ to be $f_n(V_1, \ldots, V_n, U_2)$, where $f_n : [0,1]^{n+1} \mapsto D[0,1]$ is measurable and $D[0,1]$ has the Skorohod rather than uniform metric, since both $t \mapsto \sqrt{n}(\hat{H}_n(t) - t)$ and $t \mapsto B(t)$ are Borel measurable on the Skorohod space $D[0,1]$. Since $\mathrm{P}(B \in C[0,1]) = 1$, and since the uniform and Skorohod metrics are equivalent on $C[0,1]$, we now have that $f_n$ is also measurable with respect to the uniform topology. Thus the map $g_n : \mathbb{R}^n \times [0,1] \mapsto C[0,1]$ defined by the composition $(Y_1, \ldots, Y_n, U_0) \mapsto (V_1, \ldots, V_n, U_2) \mapsto B$ is Borel measurable, and (15.31) follows.□

LEMMA 15.15 *Given two random elements $X$ and $Y$ in a separable metric space $\mathbb{X}$, there exists a Borel measurable $f : \mathbb{X} \times [0,1] \mapsto \mathbb{X}$ and a uniform random variable $Z$ independent of $X$, such that $Y = f(X, Z)$ almost surely.*

*Proof.* The result and proof are given in Skorohod (1976). While Skorohod's paper does not specify uniformity of $Z$, this readily follows without loss of generality, since any real random variable can be expressed as a right-continuous function of a uniform deviate.□

### *15.5.3 Inference for Marginal Sample Means*

For each $1 \leq j \leq p_n$, assume for this section that $F_{j(n)}$ has finite mean $\mu_{j(n)}$ and standard deviation $\sigma_{j(n)} > 0$. Let $\bar{X}_{j(n)}$ be the sample mean of $X_{1j(n)}, \ldots, X_{nj(n)}$. The following corollary yields simultaneous consistency of the marginal sample means. All proofs are given at the end of this subsection.

COROLLARY 15.16 *Assume the closure of the support of $F_{j(n)}$ is a compact interval $[a_{j(n)}, b_{j(n)}]$ with $a_{j(n)} \neq b_{j(n)}$. Under the conditions of Theorem 15.12 and with the same constant $c_0$, we have for all $n, p_n \geq 2$,*

$$\left\|\max_{1\leq j\leq p_n} |\bar{X}_{j(n)} - \mu_{j(n)}|\right\|_{\psi_2} \leq c_0\sqrt{\frac{\log p_n}{n}} \max_{1\leq j\leq p_n} |b_{j(n)} - a_{j(n)}|. \tag{15.34}$$

Note that Corollary 15.16 slightly extends the large $p$ small $n$ consistency results of van der Laan and Bryan (2001) by allowing the range of the support to increase with $n$ provided it does not increase too rapidly. Kosorok and Ma (2007) show that the bounded range assumption can be relaxed at the expense of increasing the upper bound in (15.34).

Now suppose we wish to test the marginal null hypothesis $H_0^{j(n)} : \mu_{j(n)} = \mu_{0,j(n)}$ with $T_{j(n)} = \sqrt{n}(\bar{X}_{j(n)} - \mu_{0,j(n)})/\hat{\sigma}_{j(n)}$, where $\hat{\sigma}_{j(n)}$ is a location-invariant and consistent estimator of $\sigma_{j(n)}$. To use FDR, as mentioned

previously, we need uniformly consistent estimates of the p-values of these tests. While permutation methods can be used, an easier way is to use $\hat{\pi}_{j(n)} = 2\Phi(-|T_{j(n)}|)$, where $\Phi$ is the standard normal distribution function, but we need to show this is valid. For the estimator $\hat{\sigma}_{j(n)}$, we require $\hat{\sigma}_{j(n)}/\sigma_{j(n)}$ to be uniformly consistent for 1, i.e.,

$$E_{0n} \equiv \max_{1\le j\le p_n} \left|\hat{\sigma}^2_{j(n)}/\sigma^2_{j(n)} - 1\right| = o_P(1).$$

One choice for $\hat{\sigma}^2_{j(n)}$ that satisfies this requirement is the sample variance $S^2_{j(n)}$ for $X_{1j(n)}, \ldots, X_{nj(n)}$:

COROLLARY 15.17 *Assume $n \to \infty$, with $p_n \ge 2$ and $\log p_n = o(n^\gamma)$ for some $\gamma \in (0, 1]$. Assume the closure of the support of $F_{j(n)}$ is compact as in Corollary 15.16, and let $d_n \equiv \max_{1\le j\le p_n} \sigma^{-1}_{j(n)}|b_{j(n)} - a_{j(n)}|$. Then $E_{0n} = O(n^{-1}) + o_P(d_n^2 n^{\gamma/2-1/2})$. In particular, if $d_n = O(1)$, then $E_{0n} = o_P(1)$.*

This approach leads to uniformly consistent p-values:

COROLLARY 15.18 *Assume as $n \to \infty$ that $p_n \ge 2$, $\log p_n = o(n^{1/2})$, $d_n = O(1)$ and $E_{0n} = o_P(1)$. Then there exist standard normal random variables $Z_{1(n)}, \ldots, Z_{p_n(n)}$ which are conditionally independent given $\mathcal{F}_n$, with each $Z_{j(n)}$ having conditional distribution given $\mathcal{F}_n$ depending only on $\mathcal{F}_{j(n)}$, $1 \le j \le p_n$, such that*

$$\max_{1\le j\le p_n} |\hat{\pi}_{j(n)} - \pi_{j(n)}| \quad = \quad o_P(1), \tag{15.35}$$

*where*

$$\pi_{j(n)} \quad \equiv \quad 2\Phi\left(-\left|Z_{j(n)} + \frac{\sqrt{n}(\mu_{j(n)} - \mu_{0,j(n)})}{\sigma_{j(n)}}\right|\right), \quad 1 \le j \le p_n. \tag{15.36}$$

*In particular, (15.35) holds if $\hat{\sigma}^2_{j(n)} = S^2_{j(n)}$.*

Corollary 15.18 tells us that if we assume bounded ranges of the distributions, then the approximate p-values are asymptotically valid, provided we use the sample standard deviation for t-tests and $\log p_n = o(n^{1/2})$.

*Proof of Corollary 15.16.* Apply Theorem 15.12 and the following identity:

$$\begin{aligned}\int_{[a_{j(n)},b_{j(n)}]} x\left[d\hat{F}_{j(n)}(x) - dF_{j(n)}(x)\right] \\ = \quad -\int_{[a_{j(n)},b_{j(n)}]} \left[\hat{F}_{j(n)}(x) - F_{j(n)}(x)\right] dx.\end{aligned} \tag{15.37}$$

□

*Proof of Corollary 15.17.* Let $S^2_{j(n)}$ be the sample variance version with $n$ in the denominator, and let $\tilde{S}^2_{j(n)}$ be the version with denominator $n-1$.

Then $\tilde{S}^2_{j(n)}/\sigma^2_{j(n)} - 1 = O(n^{-1}) + (1 + o(1))\left(S^2_{j(n)}/\sigma^2_{j(n)} - 1\right)$. and thus we can assume the denominator is $n$ after adding the term $O(n^{-1})$. Note that

$$\left|\frac{S^2_{j(n)}}{\sigma^2_{j(n)}} - 1\right| \le \left|n^{-1}\sum_{i=1}^{n}\frac{(X_{ij(n)} - \mu_{j(n)})^2}{\sigma^2_{j(n)}} - 1\right| + \left(\frac{\bar{X}_{j(n)} - \mu_{j(n)}}{\sigma_{j(n)}}\right)^2.$$

Apply Corollary 15.16 twice, once for the data $(X_{ij(n)} - \mu_{j(n)})^2/\sigma^2_{j(n)}$ and once for the data $(X_{ij(n)} - \mu_{j(n)})/\sigma_{j(n)}$. This gives us

$$\max_{1\le j\le p_n}\left|\frac{S^2_{j(n)}}{\sigma^2_{j(n)}} - 1\right| \le O_P\left(\sqrt{\frac{\log p_n}{n}}\, d_n^2 + \frac{\log p_n}{n}\, d_n^2\right)$$

$= o_P(d_n^2 n^{\gamma/2-1/2})$, since $n^{\gamma/2-1/2} = o(1)$ by assumption, and the desired result follows.□

*Proof of Corollary 15.18.* Note that for any $x \in \mathbb{R}$ and any $y > 0$, $|\Phi(xy) - \Phi(x)| \le 0.25 \times (|1 - y| \vee |1 - 1/y|)$. Thus

$$\begin{aligned}\max_{1\le j\le p_n}|\hat{\pi}_{j(n)} - \hat{\pi}^*_{j(n)}| &\le \frac{1}{2}\left(\max_{1\le j\le n}(\hat{\sigma}_{j(n)} \vee \sigma_{j(n)})\left|\frac{1}{\hat{\sigma}_{j(n)}} - \frac{1}{\sigma_{j(n)}}\right|\right)\\ &= O_P(E_{0n}^{1/2}),\end{aligned}$$

where $\hat{\pi}^*_{j(n)} \equiv 2\Phi(-|T^*_{j(n)}|)$ and $T^*_{j(n)} \equiv \sqrt{n}(\bar{X}_{j(n)} - \mu_{0,j(n)})/\sigma_{j(n)}$.

Now Theorem 15.13 yields

$$\begin{aligned}\max_{1\le j\le p_n}\left|\hat{\pi}^*_{j(n)} - \pi_{j(n)}\right| &= O_P\left(\frac{c_1\log n + c_2\log p_n}{\sqrt{n}}\right)\max_{1\le j\le p_n}\frac{|b_{j(n)} - a_{j(n)}|}{\sigma_{j(n)}}\\ &= o_P(n^{\gamma-1/2}d_n),\end{aligned}$$

and thus $\max_{1\le j\le p_n}|\hat{\pi}_{j(n)} - \pi_{j(n)}| = O_P(E_{0n}^{1/2}) + o_P(n^{\gamma-1/2}d_n)$. The desired results now follow.□

## 15.6 Exercises

**15.6.1.** For any $c \ge 1$, let $t \mapsto f_c(t) \equiv c^{-1}\log(1 + e^{ct})$. Show that $\sup_{t\in\mathbb{R}}|f_a(t) - f_b(t)| \le |a - b|$ and, for any $1 \le A_0 < \infty$, that

$$\sup_{a,b\ge A_0, t\in\mathbb{R}}|f_a(t) - f_b(t)| \le A_0^{-1}.$$

Use these results to show that $N(\epsilon, \mathcal{G}_2, \|\cdot\|_\infty) \le K_0\epsilon^{-2}$, for all $\epsilon > 0$.

**15.6.2.** In the proof of Theorem 15.6, it is necessary to verify that the proportional odds "model is identifiable." This means that we need to show

that whenever $(1+e^{\beta' Z}A(U))^{-1} = (1+e^{\beta_0' Z}A_0(U))^{-1}$ almost surely, $\beta = \beta_0$ and $A(t) = A_0(t)$ for all $t \in [0,\tau]$. Show that this is true. Hint: Because $A_0$ is continuous with density bounded above and below, it is enough to verify that whenever $e^{\beta' Z}a(t) = e^{\beta_0' Z}$ almost surely for every $t \in [0,\tau]$, where $a(t) \equiv dA(t)/dA_0(t)$, we have $\beta = \beta_0$ and $a(t) = 1$ for all $t \in [0,\tau]$. It may help to use the fact that $\mathrm{var}[Z]$ is positive definite.

**15.6.3.** Prove that the function $\phi$ defined in the proof of Lemma 15.7 is Lipschitz continuous on $[-k,k] \times [0,1]$ with finite Lipschitz constant depending only on $k$, for all $k < \infty$.

**15.6.4.** In the context of Section 15.3.4, verify that for any $0 < q < p < \infty$, the following are true for all $\theta \in \mathrm{lin}\Theta$:

(i) $(q/(2p))\|\theta\|_{(p)} \le \|\theta\|_{(q)} \le \|\theta\|_{(p)}$, and

(ii) $p\|\theta\|_\infty \le \|\theta\|_{(p)} \le 4p\|\theta\|_\infty$.

**15.6.5.** Verify the form of the components of $\hat{\sigma}_\theta$ given in the formulas immediately following (15.20). It may help to scrutinize the facts that $\theta(h) = \beta' h_1 + \int_0^\tau h_2(s)dA(s)$ and that the derivative of $t \mapsto PV^\tau(\theta_t)(h)$, where $\theta_t = \theta_1 + t\theta$, is $-\theta(\sigma_{\theta_1}(h))$.

**15.6.6.** For $\sigma_2$ defined in the proof of Theorem 15.9, verify that the range of $h \mapsto \sigma_2(h)$ over the unit ball in $\mathcal{H}_\infty$ lies within a compact set.

**15.6.7.** Show that (15.22) holds. Hint: Use the continuity of $\theta \mapsto V^\tau(\theta)(h)$.

**15.6.8.** In the proof of Lemma 15.11, verify that the conditional covariance of $\mathcal{B}_n$ has the form given in (15.26).

**15.6.9.** In the context of the proof of Theorem 15.14, show that there exists a function $h_n : \mathbb{R}^n \times [0,1] \mapsto [0,1]^n$ such that $(V_1,\ldots,V_n) \equiv h_n(Y_1,\ldots, Y_n, U_1)$ is a sample of i.i.d. uniforms and $(Y_1,\ldots,Y_n) = (\psi(V_1),\ldots,\psi(V_n))$, where $\psi(u) \equiv \inf\{x : G(x) \ge u\}$. You can use the result from Theorem 20.4 of Billingsley (1995) that yields the existence of a measurable function $g_m : [0,1] \mapsto [0,1]^m$, so that $(U_1',\ldots,U_m') \equiv g_m(U_1)$ are i.i.d. uniforms. The challenge is to address the possible presence of discontinuities in the distribution of $Y_1$.

## 15.7 Notes

In the setting of Section 15.4, taking the supremum of $|U_n(\zeta)|$ over $\zeta$ may not be optimal in some settings. The issue is that standard maximum likelihood optimality arguments do not hold when some of the parameters are unidentifiable under the null hypothesis. Using Bayesian arguments, Andrews and Ploberger (1994) show that it is sometimes better to compute a

weighted integral over the unidentifiable index parameter ($\zeta$ in our example) rather than taking the supremum. This issue is explored briefly in the context of transformation models in Kosorok and Song (2007).

We also note that the original motivation behind the theory developed in Section 15.5 was to find a theoretical justification for using least absolute deviation estimation for a certain semiparametric regression model for robust normalization and significance analysis of cDNA microarrays. The model, estimation, and theoretical justification for this are described in Ma, Kosorok, Huang, et al. (2006).

# Part III

# Semiparametric Inference

# 16

# Introduction to Semiparametric Inference

The goal of Part III, in combination with Chapter 3, is to provide a solid working knowledge of semiparametric inference techniques that will facilitate research in the area. Chapter 17 presents additional mathematical background, beyond that provided in Chapter 6, which enables the technical development of efficiency calculations and related mathematical tools needed for the remaining chapters. The topics covered include projections, Hilbert spaces, and adjoint operators in Banach spaces. The main topics overviewed in Chapter 3 will then be extended and generalized in Chapters 18 through 21. Many of these topics will utilize in a substantial way the empirical process methods developed in Part II. Part III, and the entire book, comes to a conclusion in the case studies presented in Chapter 22.

The overall development of the material in this last part will be logical and orderly, but not quite as linear as was the case with Part II. Only a subset of semiparametric inference topics will be covered, but the material that is covered should facilitate understanding of the omitted topics. The structure outlined in Chapter 3 should help the reader maintain perspective during the developments in the section, while the case study examples should help the reader successfully integrate the concepts discussed. Some of the concepts presented in Chapter 3 were covered sufficiently in that chapter and will not be discussed in greater detail in Part III. It might be helpful at this point for the reader to review Chapter 3 before continuing.

Many of the topics in Part III were chosen because of the author's personal interests and not because they are necessarily the most important topics in the area. In fact, there are a number of interesting and valuable topics in semiparametric inference that are omitted because of space and

time constraints. Some of these omissions will be mentioned towards the end of this chapter. In spite of these caveats, the topics included are all worthwhile and, taken together, will provide a very useful introduction for both those who wish to pursue research in semiparametric inference as well as for those who simply want to sample the area.

The main components of semiparametric efficiency theory for general models is studied in Chapter 18. The important concepts of tangent spaces, regularity and efficiency are presented with greater mathematical completeness than in Section 3.2. With the background we have from Part II of the book, this mathematical rigor adds clarity to the presentation that is helpful in research. Efficiency for parameters defined as functionals of other parameters, infinite-dimensional parameters and groups of parameters are also studied in this chapter. The delta method from Chapter 12 proves to be useful in this setting. Chapter 18 then closes with a discussion of optimality theory for hypothesis testing. As one might expect, there is a fundamental connection between efficient estimators for a parameter and optimal tests of hypotheses involving that parameter.

In Chapter 19, we focus on efficient estimation and inference for the parametric component $\theta$ in a semiparametric model that includes a potentially infinite-dimensional component $\eta$. The concepts of efficient score functions discussed in Section 3.2 and least-favorable submodels are elaborated upon and connected to profile likelihood based estimation and inference. This includes a development of some of the methods hinted at in Section 3.4 as well as some discussion of penalized likelihoods.

Estimation and inference for regular but infinite-dimensional parameters using semiparametric likelihood is developed in Chapter 20. Topics covered include joint efficient estimation of $\theta$ and $\eta$, when $\eta$ is infinite-dimensional but can still be estimated efficiently and at the $\sqrt{n}$-rate. Inference for both parameters jointly—and separately—is discussed. The usual weighted bootstrap will be discussed along with a more efficient approach, called the "piggyback bootstrap," that utilizes the profile likelihood structure. Empirical process methods from Part II will be quite useful in this chapter.

Chapter 21 will introduce and develop semiparametric estimation and inference for general semiparametric M-estimators, including non-likelihood estimation procedures such as least-squares and misspecified likelihood estimation. Because infinite-dimensional parameters are still involved, tangent spaces and other concepts from earlier chapters in Part III will prove useful in deriving limiting distributions. A weighted bootstrap procedure for inference will also be presented and studied. Entropy calculations and M-estimator theory discussed in Part II will be quite useful in these developments. The concepts, issues, and results will be illustrated with a number of interesting examples, including least-squares estimation for the Cox model with current status data and M-estimation for partly-linear logistic regression with a misspecified link function.

A number of important ideas will not be covered in these chapters, including *local efficiency*, *double robustness* and *causal inference*. The rich topic of efficiency in estimating equations will also not be discussed in detail, although the concept will be covered briefly in tandem with an example in the case studies of Chapter 22 (see Section 22.3). An estimating equation is locally efficient if it produces root-$n$ consistent estimators for a large class of models while also producing an efficient estimator for at least one of the models in that class (see Section 1.2.6 of van der Laan and Robins, 2003). A semiparametric estimation procedure is doubly robust if only part of the model needs to be correctly specified in order to achieve full efficiency for the parameter of interest (see Section 1.6 of van der Laan and Robins, 2003). Double robustness can be quite useful in missing data settings. Causal inference is inference for causal relationships between predictors and outcomes and has emerged in recent years as one of the most active and important areas of biostatistical research. Semiparametric methods are extremely useful in causal inference because of the need to develop models which flexibly adjust for confounding variables and other factors. Additional details and insights into these concepts can be found in van der Laan and Robins (2003). A somewhat less technical but very informative discussion is given in Tsiatis (2006) who also examines semiparametric missing data models in some depth.

The case studies in Chapter 22 demonstrate that empirical process methods combined with semiparametric inference tools can solve many difficult and important problems in statistical settings in a manner that makes maximum—or nearly maximum—use of the available data under both realistic and flexible assumptions. Efficiency for both finite and infinite dimensional parameters is considered. The example studied in Section 22.3 includes some discussion of general theory for estimating equations. An interesting aspect of this example is that an optimal estimating equation is not in general achievable, although improvements in efficiency are possible for certain classes of estimating equations.

As mentioned previously, many important topics are either omitted or only briefly addressed. Thus it will be important for the interested reader to pursue additional references before embarking on certain kinds of research. Another point that needs to be made is that optimality is not the only worthy goal in semiparametric inference, although it is important. An even more important aspect of semiparametric methods is their value in developing flexible and meaningful models for scientific research. This is the primary goal of this book, and it is hoped that the concepts presented will prove useful as a starting point for research in empirical processes and semiparametric inference.

# 17

# Preliminaries for Semiparametric Inference

In this chapter, we present additional technical background, beyond that given in Chapter 6, needed for the development of semiparametric inference theory. We first discuss projections without direct reference to Hilbert spaces. Next, we present basic properties and results for Hilbert spaces and revisit projections as objects in Hilbert spaces. Finally, we build on the Banach space material presented in Chapter 6, and in other places such as Chapter 12. Concepts discussed include adjoints of operators and dual spaces. The background in this chapter provides the basic framework for a "semiparametric efficiency calculus" that is used for important computations in semiparametric inference research. The development of this calculus will be the main goal of the next several chapters.

## 17.1 Projections

Geometrically, the projection of an object $T$ onto a space $\mathcal{S}$ is the element $\hat{S} \in \mathcal{S}$ that is "closest" to $T$, provided such an element exists. In the semiparametric inference context, the object is usually a random variable and the spaces of interest for projection are usually sets of square-integrable random variables. The following theorem gives a simple method for identifying the projection in this setting:

THEOREM 17.1 *Let $\mathcal{S}$ be a linear space of real random variables with finite second moments. Then $\hat{S}$ is the projection of $T$ onto $\mathcal{S}$ if and only if*

*(i) $\hat{S} \in S$ and*

*(ii)* $E(T - \hat{S})S = 0$ *for all* $S \in \mathcal{S}$.

*If* $S_1$ *and* $S_2$ *are both projections, then* $S_1 = S_2$ *almost surely. If* $\mathcal{S}$ *contains the constants, then* $ET = E\hat{S}$ *and* $cov(T - \hat{S}, S) = 0$ *for all* $S \in \mathcal{S}$.

**Proof.** First assume (i) and (ii) hold. Then, for any $S \in \mathcal{S}$, we have

$$E(T - S)^2 = E(T - \hat{S})^2 + 2E(T - \hat{S})(\hat{S} - S) + E(\hat{S} - S)^2. \quad (17.1)$$

But Conditions (i) and (ii) force the middle term to be zero, and thus $E(T - S)^2 \geq E(T - \hat{S})^2$ with strict inequality whenever $E(\hat{S} - S)^2 > 0$. Thus $\hat{S}$ is the almost surely unique projection of $T$ onto $\mathcal{S}$.

Conversely, assume that $\hat{S}$ is a projection and note that for any $\alpha \in \mathbb{R}$ and any $S \in \mathcal{S}$,

$$E(T - \hat{S} - \alpha S)^2 - E(T - \hat{S})^2 = -2\alpha E(T - \hat{S})S + \alpha^2 ES^2.$$

Since $\hat{S}$ is a projection, the left side is strictly nonnegative for every $\alpha$. But the parabola $\alpha \mapsto \alpha^2 ES^2 - 2\alpha E(T - \hat{S})S$ is nonnegative for all $\alpha$ and $S$ only if $E(T - \hat{S})S = 0$ for all $S$. Thus (ii) holds, and (i) is part of the definition of a projection and so holds automatically.

The uniqueness follows from application of (17.1) to both $S_1$ and $S_2$, forcing $E(S_1 - S_2)^2 = 0$. If the constants are in $\mathcal{S}$, then Condition (ii) implies that $E(T - \hat{S})c = 0$ for any $c \in \mathbb{R}$ including $c = 1$. Thus the final assertion of the theorem also follows.□

Note that the theorem does not imply that a projection always exists. In fact, if the set $\mathcal{S}$ is open in the $L_2(P)$ norm, then the infimum of $E(T-S)^2$ over $S \in \mathcal{S}$ is not achieved. A sufficient condition for existence, then, is that $\mathcal{S}$ be closed in the $L_2(P)$ norm, but often existence can be established directly. We will visit this issue more in the upcoming chapters.

A very useful example of a projection is a conditional expectation. Let $X$ and $Y$ be real random variables on a probability space. Then $g_0(y) \equiv E(X|Y = y)$ is the conditional expectation of $X$ given $Y = y$. If we let $\mathcal{G}$ be the space of all measurable functions $g$ of $Y$ such that $Eg^2(Y) < \infty$, then it is easy to verify that

$$E(X - g_0(Y))g(Y) = 0, \text{ for every } g \in \mathcal{G}.$$

Thus, provided $Eg_0^2(Y) < \infty$, $E(X|Y = y)$ is the projection of $X$ onto the linear space $\mathcal{G}$. By Theorem 17.1, the conditional expectation is almost surely unique. We will utilize conditional expectations frequently for calculating the projections needed in semiparametric inference settings.

## 17.2 Hilbert Spaces

A Hilbert space is essentially an abstract generalization of a finite-dimensional Euclidean space. This abstraction is a special case of a Banach space

and, like a Banach space, is often infinite-dimensional. To be precise, a Hilbert space is a Banach space with an inner product. An inner product on a Banach space $\mathbb{D}$ with norm $\|\cdot\|$ is a function $\langle\cdot,\cdot\rangle : \mathbb{D} \times \mathbb{D} \mapsto \mathbb{R}$ such that, for all $\alpha, \beta \in \mathbb{R}$ and $x, y, z \in \mathbb{D}$, the following hold:

(i) $\langle x, x\rangle = \|x\|^2$,

(ii) $\langle x, y\rangle = \langle y, x\rangle$, and

(iii) $\langle \alpha x + \beta y, z\rangle = \alpha\langle x, z\rangle + \beta\langle y, z\rangle$.

A semi-inner product arises when $\|\cdot\|$ is a semi-norm.

It is also possible to generate a norm beginning with an inner product. Let $\mathbb{D}$ be a linear space with semi-inner product $\langle\cdot,\cdot\rangle$ (not necessarily with a norm). In other words, $\langle\cdot,\cdot\rangle$ satisfies Conditions (ii)–(iii) of the previous paragraph and also satisfies $\langle x, x\rangle \geq 0$. Then $\|x\| \equiv \langle x, x\rangle^{1/2}$, for all $x \in \mathbb{D}$, defines a semi-norm. This is verified in the following theorem:

THEOREM 17.2 *Let $\langle\cdot,\cdot\rangle$ be a semi-inner product on $\mathbb{D}$, with $\|x\| \equiv \langle x, x\rangle^{1/2}$ for all $x \in \mathbb{D}$. Then, for all $\alpha \in \mathbb{R}$ and all $x, y \in \mathbb{D}$,*

*(a)* $\langle x, y\rangle \leq \|x\|\,\|y\|$,

*(b)* $\|x + y\| \leq \|x\| + \|y\|$, *and*

*(c)* $\|\alpha x\| = |\alpha| \times \|x\|$.

*Moreover, if $\langle\cdot,\cdot\rangle$ is also an inner product, then $\|x\| = 0$ if and only if $x = 0$.*

**Proof.** Note that

$$0 \leq \langle x - \alpha y, x - \alpha y\rangle = \langle x, x\rangle - 2\alpha\langle x, y\rangle + \alpha^2\langle y, y\rangle,$$

from which we can deduce

$$0 \leq at^2 + bt + c \equiv q(t),$$

where $c \equiv \langle x, x\rangle$, $b \equiv -2|\langle x, y\rangle|$, $a \equiv \langle y, y\rangle$, and $t \equiv \alpha\,\mathrm{sign}\langle x, y\rangle$ (which is clearly free to vary over $\mathbb{R}$). This now forces $q(t)$ to have at most one real root. This implies that the discriminant from the quadratic formula is not positive. Hence

$$0 \geq b^2 - 4ac = 4\langle x, y\rangle^2 - 4\langle x, x\rangle\langle y, y\rangle,$$

and Part (a) follows.

Using (a), we have for any $x, y \in \mathbb{D}$,

$$\begin{aligned} \|x + y\|^2 &= \langle x + y, x + y\rangle \\ &= \|x\|^2 + 2\langle x, y\rangle + \|y\|^2 \\ &\leq \|x\|^2 + 2\|x\|\,\|y\| + \|y\|^2 \\ &= (\|x\| + \|y\|)^2, \end{aligned}$$

and (b) follows. Part (c) follows readily from the equalities

$$\|\alpha x\|^2 = \langle \alpha x, \alpha x \rangle = \alpha \langle \alpha x, x \rangle = \alpha^2 \langle x, x \rangle = (\alpha \|x\|)^2.$$

Part (d) follows readily from the definitions.□

Part (a) in Theorem 17.2 is also known as the Cauchy-Schwartz inequality. Two elements $x, y$ in a Hilbert space $\mathbb{H}$ with inner product $\langle \cdot, \cdot \rangle$ are *orthogonal* if $\langle x, y \rangle = 0$, denoted $x \perp y$. For any set $C \subset \mathbb{H}$ and any $x \in \mathbb{H}$, $x$ is orthogonal to $C$ if $x \perp y$ for every $y \in C$, denoted $x \perp C$. The two subsets $C_1, C_2 \subset \mathbb{H}$ are orthogonal if $x \perp y$ for every $x \in C_1$ and $y \in C_2$, denoted $C_1 \perp C_2$. For any $C_1 \subset \mathbb{H}$, the *orthocomplement* of $C_1$, denoted $C_1^{\perp}$, is the set $\{x \in \mathbb{H} : x \perp C_1\}$.

Let the subspace $H \subset \mathbb{H}$ be linear and closed. By Theorem 17.1, we have for any $x \in \mathbb{H}$ that there exists an element $y \in H$ that satisfies $\|x - y\| \leq \|x - z\|$ for any $z \in H$, and such that $\langle x - y, z \rangle = 0$ for all $z \in H$. Let $\Pi$ denote an operator that performs this projection, i.e., let $\Pi x \equiv y$ where $y$ is the projection of $x$ onto $H$. The "projection" operator $\Pi : \mathbb{H} \mapsto H$ has several important properties which we give in the following theorem. Recall from Section 6.3 the definitions of the null space $N(T)$ and range $R(T)$ of an operator $T$.

THEOREM 17.3 *Let $H$ be a closed, linear subspace of $\mathbb{H}$, and let $\Pi : \mathbb{H} \mapsto H$ be the projection operator onto $H$. Then*

*(i)* $\Pi$ *is continuous and linear,*

*(ii)* $\|\Pi x\| \leq \|x\|$ *for all* $x \in \mathbb{H}$,

*(iii)* $\Pi^2 \equiv \Pi \Pi = \Pi$, *and*

*(iv)* $N(\Pi) = H^{\perp}$ *and* $R(\Pi) = H$.

**Proof.** Let $x, y \in \mathbb{H}$ and $\alpha, \beta \in \mathbb{R}$. If $z \in H$, then

$$\begin{aligned} \langle [\alpha x + \beta y] - [\alpha \Pi x + \beta \Pi y], z \rangle &= \alpha \langle x - \Pi x, z \rangle + \beta \langle y - \Pi y, z \rangle \\ &= 0. \end{aligned}$$

Now by the uniqueness of projections (via Theorem 17.1), we now have that $\Pi(\alpha x + \beta y) = \alpha \Pi x + \beta \Pi y$. This yields linearity of $\Pi$. If we can establish (ii), (i) will follow. Since $\langle x - \Pi x, \Pi x \rangle = 0$, for any $x \in \mathbb{H}$, we have

$$\|x\|^2 = \|x - \Pi x\|^2 + \|\Pi x\|^2 \geq \|\Pi x\|^2,$$

and (ii) (and hence also (i)) follows.

For any $y \in H$, $\Pi y = y$. Thus for any $x \in \mathbb{H}$, $\Pi(\Pi x) = \Pi x$, and (iii) follows. Now let $x \in N(\Pi)$. Then $x = x - \Pi x \in H^{\perp}$. Conversely, for any $x \in H^{\perp}$, $\Pi x = 0$ by definition, and thus $x \in N(\Pi)$. Hence $N(\Pi) = H^{\perp}$. Now it is trivial to verify that $R(\Pi) \subset H$ by the definitions. Moreover, for any $x \in H$, $\Pi x = x$, and thus $H \subset R(\Pi)$. This implies (iv).□

One other point we note is that for any projection $\Pi$ onto a closed linear subspace $H \subset \mathbb{H}$, $I - \Pi$, where $I$ is the identity, is also a projection onto the closed linear subspace $H^{\perp}$. An important example of a Hilbert space $\mathbb{H}$, one that will be of great interest to us throughout this chapter, is $\mathbb{H} = L_2(P)$ with inner product $\langle f, g \rangle = \int f g dP$. A closed, linear subspace of interest to us is $L_2^0(P) \subset L_2(P)$ which consists of all mean zero functions in $L_2(P)$. The projection operator $\Pi : L_2(P) \mapsto L_2^0(P)$ is $\Pi x = x - Px$. To see this, note that $\Pi x \in L_2^0(P)$ and $\langle x - \Pi x, y \rangle = \langle Px, y \rangle = Px\, Py = 0$ for all $y \in L_2^0(P)$. Thus by Theorem 17.1, $\Pi x$ is the unique projection onto $L_2^0(P)$. It is also not hard to verify that $I - \Pi$ is the projection onto the constants (see Exercise 17.4.2).

Now consider the situation where there are two closed, linear subspaces $H_1, H_2 \subset \mathbb{H}$ but $H_1$ and $H_2$ are not necessarily orthogonal. Let $\Pi_j$ be the projection onto $H_j$, and define $Q_j \equiv I - \Pi_j$, for $j = 1, 2$. The idea of *alternating projections* is to alternate between projecting $h \in \mathbb{H}$ onto the orthocomplements of $H_1$ and $H_2$ repeatedly, so that in the limit we obtain the projection $\tilde{h}$ of $h$ onto the orthocomplement of the closure of the *sumspace* of $H_1$ and $H_2$, denoted $\overline{H_1 + H_2}$. The sumspace of $H_1$ and $H_2$ is simply the closed linear span of $\{h_1 + h_2 : h_1 \in H_1, h_2 \in H_2\}$. Then $h - \tilde{h}$ is hopefully the projection of $h$ onto $\overline{H_1 + H_2}$.

For any $h \in \mathbb{H}$, let $h_j^{(m)} \equiv \Pi_j[I - (Q_1 Q_2)^m]h$ and $\tilde{h}_j^{(m)} \equiv \Pi_j[I - (Q_2 Q_1)^m]h$, for $j = 1, 2$. Also, let $\Pi$ be the projection onto $\overline{H_1 + H_2}$ and $Q \equiv I - \Pi$. The results of the following theorem are that $\Pi$ can be computed as the limit of iterations between $Q_2$ and $Q_1$, and that $\Pi h$ can be expressed as a sum of elements in $H_1$ and $H_2$ under certain conditions. These results will be useful in the partly linear Cox model case study of Chapter 22. We omit the proof which can be found in Section A.4 of Bickel, Klaassen, Ritov and Wellner (1998) (hereafter abbreviated BKRW):

THEOREM 17.4 *Assume that $H_1, H_2 \subset \mathbb{H}$ are closed and linear. Then, for any $h \in \mathbb{H}$,*

*(i)* $\|h_1^{(m)} + h_2^{(m)} - \Pi h\| \equiv u_m \to 0$, *as* $m \to \infty$;

*(ii)* $\|\tilde{h}_1^{(m)} + \tilde{h}_2^{(m)} - \Pi h\| \equiv \tilde{u}_m \to 0$, *as* $m \to \infty$;

*(iii)* $u_m \vee \tilde{u}_m \leq \rho^{2(m-1)}$, *where $\rho$ is the cosine of the minimum angle $\tau$ between $H_1$ and $H_2$ considering only elements in $(H_1 \cap H_2)^{\perp}$;*

*(iv)* $\rho < 1$ *if and only if $H_1 + H_2$ is closed; and*

*(v)* $\|I - (Q_2 Q_1)^m - \Pi\| = \|I - (Q_1 Q_2)^m - \Pi\| = \rho^{2m-1}$.

We close this section with a brief discussion of *linear functionals* on Hilbert spaces. Recall from Section 6.3 the definition of a linear operator and the fact that the norm for a linear operator $T : \mathbb{D} \mapsto \mathbb{E}$ is $\|T\| \equiv \sup_{x \in \mathbb{D}: \|x\| \leq 1} \|Tx\|$. In the special case where $\mathbb{E} = \mathbb{R}$, a linear operator is

called a linear functional. A linear functional $T$, like a linear operator, is bounded when $\|T\| < \infty$. By Proposition 6.15, boundedness is equivalent to continuity in this setting. A very important result for bounded linear functionals in Hilbert spaces is the following:

THEOREM 17.5 *(**Riesz representation theorem**) If $L : \mathbb{H} \mapsto \mathbb{R}$ is a bounded linear functional on a Hilbert space, then there exists a unique element $h_0 \in \mathbb{H}$ such that $L(h) = \langle h, h_0 \rangle$ for all $h \in \mathbb{H}$, and, moreover, $\|L\| = \|h_0\|$.*

**Proof.** Let $H = N(L)$, and note that $H$ is closed and linear because $L$ is continuous and the space $\{0\}$ (consisting of only zero) is trivially closed and linear. Assume that $H \neq \mathbb{H}$ since otherwise the proof would be trivial with $h_0 = 0$. Thus there is a vector $f_0 \in H^\perp$ such that $L(f_0) = 1$. Hence, for all $h \in \mathbb{H}$, $h - L(h)f_0 \in H$, since

$$L(h - L(h)f_0) = L(h) - L(h)L(f_0) = 0.$$

Thus, by the orthogonality of $H$ and $H^\perp$, we have for all $h \in \mathbb{H}$ that

$$0 = \langle h - L(h)f_0, f_0 \rangle = \langle h, f_0 \rangle - L(h)\|f_0\|^2.$$

Hence if $h_0 \equiv \|f_0\|^{-2} f_0$, $L(h) = \langle h, h_0 \rangle$ for all $h \in \mathbb{H}$.

Suppose $h_0' \in \mathbb{H}$ satisfies $\langle h, h_0' \rangle = \langle h, h_0 \rangle$ for all $h \in \mathbb{H}$, then $h_0 - h_0' \perp \mathbb{H}$, and thus $h_0' = h_0$. Since by the Cauchy-Schwartz inequality $|\langle h, h_0 \rangle| \leq \|h\| \, \|h_0\|$ and also $\langle h_0, h_0 \rangle = \|h_0\|^2$, we have $\|L\| = \|h_0\|$, and thus the theorem is proved.□

## 17.3 More on Banach Spaces

Recall the discussion about Banach spaces in Section 6.1. As with Hilbert spaces, a linear functional on a Banach space is just a linear operator with real range. The *dual space* $\mathbb{B}^*$ of a Banach space $\mathbb{B}$ is the set of all continuous, linear functionals on $\mathbb{B}$. By applying Proposition 6.15, it is clear that every $b^* \in \mathbb{B}^*$ satisfies $|b^* b| \leq \|b^*\| \, \|b\|$ for every $b \in \mathbb{B}$, where $\|b^*\| \equiv \sup_{b \in \mathbb{B}: \|b\| \leq 1} |b^* b| < \infty$.

For the special case of a Hilbert space $\mathbb{H}$, $\mathbb{H}^*$ can identified with $\mathbb{H}$ by the Reisz representation theorem given above. This implies that there exists an *isometry* (a one-to-one map that preserves norms) between $\mathbb{H}$ and $\mathbb{H}^*$. To see this, choose $h^* \in \mathbb{H}^*$ and let $\tilde{h} \in \mathbb{H}$ be the unique element that satisfies $\langle h, \tilde{h} \rangle = h^* h$ for all $h \in \mathbb{H}$. Then

$$\|h^*\| = \sup_{h \in \mathbb{H}: \|h\| \leq 1} |\langle h, \tilde{h} \rangle| \leq \|\tilde{h}\|$$

by the Cauchy-Schwartz inequality. The desired conclusion follows since $h^*$ was arbitrary.

We now return to the generality of Banach spaces. For each continuous, linear operator between Banach spaces $A : \mathbb{B}_1 \mapsto \mathbb{B}_2$, there exists an *adjoint map* (or just adjoint) $A^* : \mathbb{B}_2^* \mapsto \mathbb{B}_1^*$ defined by $(A^* b_2^*)b_1 = b_2^* A b_1$ for all $b_1 \in \mathbb{B}_1$ and $b_2^* \in \mathbb{B}_2^*$. It is straightforward to verify that the resulting $A^*$ is linear. The following proposition tells us that $A^*$ is also continuous (by being bounded):

PROPOSITION 17.6 *Let $A : \mathbb{B}_1 \mapsto \mathbb{B}_2$ be a bounded linear operator between Banach spaces. Then $\|A^*\| = \|A\|$.*

**Proof.** Since also, for any $b_2^* \in \mathbb{B}_2^*$,

$$\begin{aligned} \|A^* b_2^*\| &= \sup_{b_1 \in \mathbb{B}_1 : \|b_1\| \leq 1} |A^* b_2^* b_1| \\ &= \sup_{b_1 \in \mathbb{B}_1 : \|b_1\| \leq 1} \left\{ \left| b_2^* \left( \frac{Ab_1}{\|Ab_1\|} \right) \right| \, \|Ab_1\| \right\} \\ &\leq \|b_2^*\| \, \|A\|, \end{aligned}$$

we have $\|A^*\| \leq \|A\|$. Thus $\|A^*\|$ is a continuous, linear operator.

Now let $A^{**} : \mathbb{B}_1^{**} \mapsto \mathbb{B}_2^{**}$ be the adjoint of $A^*$ with respect to the double duals (duals of the duals) of $\mathbb{B}_1$ and $\mathbb{B}_2$. Note that for $j = 1, 2$, $\mathbb{B}_j \subset \mathbb{B}_j^{**}$, since for any $b_j \in \mathbb{B}_j$, the map $b_j : \mathbb{B}_j^* \mapsto \mathbb{R}$ defined by $b_j^* \mapsto b_j^* b_j$, is a bounded linear functional. By the definitions involved, we now have for any $b_1 \in \mathbb{B}_1$ and $b_2^* \in \mathbb{B}_2^*$ that

$$(A^{**} b_1) b_2^* = (A^* b_2^*) b_1 = b_2^* A b_1,$$

and thus $\|A^{**}\| \leq \|A^*\|$ and the restriction of $A^{**}$ to $\mathbb{B}_1$, denoted hereafter $A_1^{**}$, equals $A$. Hence $\|A\| = \|A_1^{**}\| \leq \|A^*\|$, and the desired result follows.□

We can readily see that the adjoint of an operator $A : \mathbb{H}_1 \mapsto \mathbb{H}_2$ between two Hilbert spaces, with respective inner products $\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$, is a map $A^* : \mathbb{H}_2 \mapsto \mathbb{H}_1$ satisfying $\langle Ah_1, h_2 \rangle_2 = \langle h_1, A^* h_2 \rangle_1$ for every $h_1 \in \mathbb{H}_1$ and $h_2 \in \mathbb{H}_2$. Note that we are using here the isometry for Hilbert spaces described at the beginning of this section. Now consider the adjoint of a restriction of a continuous linear operator $A : \mathbb{H}_1 \mapsto \mathbb{H}_2$, $A_0 : \mathbb{H}_{0,1} \subset \mathbb{H}_1 \mapsto \mathbb{H}_2$, where $\mathbb{H}_{0,1}$ is a closed, linear subspace of $\mathbb{H}_1$. If $\Pi : \mathbb{H}_1 \mapsto \mathbb{H}_{0,1}$ is the projection onto the subspace, it is not hard to verify that the adjoint of $A_0$ is $A_0^* \equiv \Pi \circ A^*$ (see Exercise 17.4.3).

Recall from Section 6.3 the notation $B(\mathbb{D}, \mathbb{E})$ denoting the collection of all linear operators between normed spaces $\mathbb{D}$ and $\mathbb{E}$. From Lemma 6.16, we know that for a given $T \in B(\mathbb{B}_1, \mathbb{B}_2)$, for Banach spaces $\mathbb{B}_1$ and $\mathbb{B}_2$, $R(T)$ is not closed unless $T$ is continuously invertible. We now give an interesting counter-example that illustrates this. Let $\mathbb{B}_1 = \mathbb{B}_2 = L_2(0, 1)$, and define $T : L_2(0, 1) \mapsto L_2(0, 1)$ by $Tx(u) \equiv ux(u)$. Then $\|T\| \leq 1$, and thus $T \in B(L_2(0, 1), L_2(0, 1))$. However, it is clear that

$$R(T) = \left\{ y \in L_2(0, 1) : \int_0^1 u^{-2} y^2(u) du < \infty \right\}. \tag{17.2}$$

Although $R(T)$ is dense in $L_2(0,1)$ (see Exercise 17.4.3), the functions $y_1(u) \equiv 1$ and $y_2(u) \equiv \sqrt{u}$ are clearly not in $R(T)$. Thus $R(T)$ is not closed.

This lack of closure of $R(T)$ arises from the simple fact that the inverse operator $T^{-1}y(u) = u^{-1}y(u)$ is not bounded over $y \in L_2(0,1)$ (consider $y = 1$, for example). On the other hand, it is easy to verify that for any normed spaces $\mathbb{D}$ and $\mathbb{E}$ and any $T \in B(\mathbb{D}, \mathbb{E})$, $N(T)$ is always closed as a direct consequence of the continuity of $T$. Observe also that for any $T \in B(\mathbb{B}_1, \mathbb{B}_2)$,

$$\begin{aligned} N(T^*) &= \{b_2^* \in \mathbb{B}_2^* : (T^* b_2^*) b_1 = 0 \text{ for all } b_1 \in \mathbb{B}_1\} \\ &= \{b_2^* \in \mathbb{B}_2^* : b_2^*(T b_1) = 0 \text{ for all } b_1 \in \mathbb{B}_1\} \\ &= R(T)^{\perp}, \end{aligned} \tag{17.3}$$

where $R(T)^{\perp}$ is an abuse of notation used in this context to denote the linear functionals in $\mathbb{B}_2^*$ that yield zero on $R(T)$. For Hilbert spaces, the notation is valid because of the isometry between a Hilbert space $\mathbb{H}$ and its dual $\mathbb{H}^*$. The identity (17.3) has the following interesting extension:

THEOREM 17.7 *For two Banach spaces $\mathbb{B}_1$ and $\mathbb{B}_2$ and for any $T \in B(\mathbb{B}_1, \mathbb{B}_2)$, $R(T) = \mathbb{B}_2$ if and only if $N(T^*) = \{0\}$ and $R(T^*)$ is closed.*

**Proof.** If $R(T) = \mathbb{B}_2$, then $0 = R(T)^{\perp} = N(T^*)$ by (17.3), and thus $T^*$ is one-to-one. Moreover, since $\mathbb{B}_2$ is closed, we also have that $R(T^*)$ is closed by Lemma 17.8 below. Conversely, if $N(T^*) = \{0\}$ and $R(T^*)$ is closed, then, by reapplication of (17.3), $R(T)^{\perp} = \{0\}$ and $R(T)$ is closed, which implies $R(T) = \mathbb{B}_2$.□

LEMMA 17.8 *For two Banach spaces $\mathbb{B}_1$ and $\mathbb{B}_2$ and for any $T \in B(\mathbb{B}_1, \mathbb{B}_2)$, $R(T)$ is closed if and only if $R(T^*)$ is closed.*

**Proof.** This is part of Theorem 4.14 of Rudin (1991), and we omit the proof.□

If we specialize (17.3) to Hilbert spaces $\mathbb{H}_1$ and $\mathbb{H}_2$, we obtain trivially for any $A \in B(\mathbb{H}_1, \mathbb{H}_2)$ that $R(A)^{\perp} = N(A^*)$. We close this section with the following useful, additional result for Hilbert spaces:

THEOREM 17.9 *For two Hilbert spaces $\mathbb{H}_1$ and $\mathbb{H}_2$ and any $A \in B(\mathbb{H}_1, \mathbb{H}_2)$, $R(A)$ is closed if and only if $R(A^*)$ is closed if and only if $R(A^*A)$ is closed. Moreover, if $R(A)$ is closed, then $R(A^*) = R(A^*A)$ and*

$$A(A^*A)^{-1}A^* : \mathbb{H}_2 \mapsto \mathbb{H}_2$$

*is the projection onto $R(A)$.*

**Proof.** The result that $R(A)$ is closed if and only if $R(A^*)$ holds by Lemma 17.8, but we will prove this again for the specialization to Hilbert

spaces to highlight some interesting features that will be useful later in the proof.

First assume $R(A)$ is closed, and let $A_0^*$ be the restriction of $A^*$ to $R(A)$, and note that $R(A^*) = R(A_0^*)$, since $R(A)^\perp = N(A^*)$ according to (17.3). Also let $A_0$ be the restriction of $A$ to $N(A)^\perp$, and note that $R(A_0) = R(A)$ by definition of $N(A)$ and therefore $R(A_0)$ is closed with $N(A_0) = \{0\}$. Thus by Part (ii) of Lemma 6.16, there exits a $c > 0$ such that $\|A_0 x\| \geq c\|x\|$ for all $x \in N(A)^\perp$. Now fix $y \in R(A_0)$, and note that there exists an $x \in N(A)^\perp$ such that $y = A_0 x$. Thus

$$\|x\| \, \|A_0^* y\| \geq \langle x, A_0^* y \rangle = \langle A_0 x, y \rangle = \|A_0 x\| \, \|y\| \geq c\|x\| \, \|y\|,$$

and therefore $\|A_0^* y\| \geq c\|y\|$ for all $y \in R(A_0)$. This means by reapplication of Part (ii) of Lemma 6.16 that $R(A^*) = R(A_0^*)$ is closed.

Now assume $R(A^*)$ is closed. By identical arguments to those used in the previous paragraph, we know that $R(A^{**})$ is closed. But because we are restricted to Hilbert spaces, $A^{**} = A$, and thus $R(A)$ is closed. Now assume that either $R(A)$ or $R(A^*)$ is closed. By what we have proven so far, we now know that both $R(A)$ and $R(A^*)$ must be closed. By recycling arguments, we know that $R(A^*A) = R(A_0^* A_0)$. Now, by applying yet again Part (ii) of Lemma 6.16, we have that there exists $c_1, c_2 > 0$ such that $\|A_0 x\| \geq c_1 \|x\|$ and $\|A_0^* y\| \geq c_2 \|y\|$ for all $x \in N(A)^\perp$ and $y \in R(A)$. Thus for all $x \in N(A)^\perp$,

$$\|A_0^* A_0 x\| \geq c_2 \|A_0 x\| \geq c_1 c_2 \|x\|,$$

and thus by both parts of Lemma 6.16, $R(A^*A) = R(A_0^* A_0)$ is closed.

Now assume that $R(A^*A)$ is closed. Thus $R(A_0^* A_0)$ is also closed and, by recycling arguments, we have that $N(A_0^* A_0) = N(A_0) = \{0\}$. Thus there exists a $c > 0$ such that for all $x \in N(A_0)^\perp$,

$$c\|x\| \leq \|A_0^* A_0 x\| \leq \|A_0^*\| \, \|A_0 x\|,$$

and therefore $R(A) = R(A_0)$ is closed. This completes the first part of the proof.

Now assume that $R(A)$ is closed and hence so also is $R(A^*)$ and $R(A^*A)$. Clearly $R(A^*A) \subset R(A^*)$. Since $R(A)^\perp = N(A^*)$, we also have that $R(A^*) \subset R(A^*A)$, and thus $R(A^*) = R(A^*A)$. By Part (i) of Lemma 6.16, we now have that $A^*A$ is continuously invertible on $R(A^*)$, and thus $A(A^*A)^{-1}A^*$ is well defined on $\mathbb{H}_2$. Observe that for any $y \in R(A)$, there exists an $x \in \mathbb{H}_1$ such that $y = Ax$, and thus

$$\Pi y \equiv A(A^*A)^{-1}A^* y = A(A^*A)^{-1}(A^*A)x = Ax = y.$$

For any $y \in R(A)^\perp$, we have $y \in N(A^*)$ and thus $\Pi y = 0$. Hence $\Pi$ is the projection operator onto $R(A)$, and the proof is complete.□

## 17.4 Exercises

**17.4.1.** In the context of Theorem 17.3, show that for any projection $\Pi : \mathbb{H} \mapsto H$, where $H$ is closed and linear, $I - \Pi$ is also a projection onto $H^{\perp}$.

**17.4.2.** In the projection example given in the paragraph following the proof of Theorem 17.3, show that $I - \Pi$ is the projection onto the space of constants, where $\Pi$ is the projection from $L_2(P)$ to $L_2^0(P)$.

**17.4.3.** Let $\mathbb{H}_1$ and $\mathbb{H}_2$ be two Hilbert spaces and let $A_0 : \mathbb{H}_{0,1}$ be the restriction to the closed subspace $\mathbb{H}_{0,1}$ of a continuous linear operator $A : \mathbb{H}_1 \mapsto \mathbb{H}_2$. If $\Pi : \mathbb{H}_1 \mapsto \mathbb{H}_{0,1}$ is the projection onto $\mathbb{H}_{0,1}$, show that the adjoint of $A_0$ is $\Pi \circ A^*$.

**17.4.4.** Show that $R(T)$ defined in (17.2) is dense in $L_2(0,1)$.

**17.4.5.** In the last paragraph of the proof of Theorem 17.9, explain why $R(A)^{\perp} = N(A^*)$ implies $R(A^*) \subset R(A^*A)$. Hint: Divide the domain of $A^*$ into two components, $R(A)$ and $R(A)^{\perp}$.

## 17.5 Notes

Much of the presentation in the first section on projections was inspired by Sections 11.1 and 11.2 of van der Vaart (1998), and Theorem 17.1 is van der Vaart's Theorem 11.1. Parts of Sections 17.2 and 17.3 were inspired by material in Section 25.2 of van der Vaart (1998), although much of technical content of Section 17.2 came from Conway (1990) and Appendix 4 of BKRW, while much of the technical content of Section 17.3 came from Chapter 4 of Rudin (1991) and Appendix 1 of BKRW.

Theorem 17.2 is a composition of Conway's Inequality 1.4 and Corollary 1.5, while Theorems 17.3 and 17.5 are modifications of Conway's Theorems 2.7 and 3.4, respectively. Theorem 17.4 is a minor modification of Theorem A.4.2 of BKRW: Parts (i), (ii) and (iii) correspond to Parts A and B of the theorem in BKRW, while Parts (iv) and (v) correspond to Parts C and D of the theorem in BKRW.

Proposition 17.6 is part of Theorem 4.10 of Rudin, and Theorem 17.7 is Theorem 4.15 of Rudin. The counter-example in $L_2(0,1)$ of the operator $Tx(u) \equiv ux(u)$ is Example A.1.11 of BKRW. The basic content of Theorem 17.9 was inspired by the last few sentences of van der Vaart's Section 25.2, although the proof is new.

# 18
# Semiparametric Models and Efficiency

The goal of this chapter is to expand in a rigorous fashion on the concepts of statistical models and efficiency introduced in Section 3.1, and it would be a good idea at this point for the reader to recall and briefly review the content of that section. Many of the results in this chapter are quite general and are not restricted to semiparametric models per se.

We begin with an in-depth discussion of the relationship between tangent sets and the concept of regularity. This includes several characterizations of regularity for general Banach-valued parameters. We then present several important results that characterize asymptotic efficiency along with several useful tools for establishing asymptotic efficiency for both finite and infinite-dimensional parameters. We conclude with a discussion of optimal hypothesis testing for one-dimensional parameters.

## 18.1 Tangent Sets and Regularity

For a statistical model $\{P \in \mathcal{P}\}$ on a sample space $\mathcal{X}$, a one-dimensional model $\{P_t\}$ is a *smooth submodel at* $P$ if $P_0 = P$, $\{P_t : t \in N_\epsilon\} \subset \mathcal{P}$ for some $\epsilon > 0$, and (3.1) holds for some measurable "tangent" function $g : \mathcal{X} \mapsto \mathbb{R}$. Here, $N_\epsilon$ is either the set $[0, \epsilon)$ (as was the case throughout Chapter 3) or $(-\epsilon, \epsilon)$ (as will be the case hereafter). Also, $P$ is usually the true but unknown distribution of the data. Note that Lemma 11.11 forces the $g$ in (3.1) to be contained in $L_2^0(P)$.

A tangent set $\dot{\mathcal{Q}}_P \subset L_2^0(P)$ *represents* a submodel $\mathcal{Q} \subset \mathcal{P}$ at $P$ if the following hold:

(i) For every smooth one-dimensional submodel $\{P_t\}$ for which

$$P_0 = P \quad \text{and} \quad \{P_t : t \in N_\epsilon\} \subset \mathcal{Q} \text{ for some } \epsilon > 0, \tag{18.1}$$

and for which (3.1) holds for some $g \in L_2^0(P)$, we have $g \in \dot{\mathcal{Q}}_P$; and

(ii) For every $g \in \dot{\mathcal{Q}}_P$, there exists a smooth one-dimensional submodel $\{P_t\}$ such that (18.1) and (3.1) both hold.

An appropriate question to ask at this point is why the focus on one-dimensional submodels? The basic reason is that score functions for finite dimensional submodels can be represented by tangent sets corresponding to one-dimensional submodels. To see this, let $\mathcal{Q} \equiv \{P_\theta : \theta \in \Theta\} \subset \mathcal{P}$, where $\Theta \subset \mathbb{R}^k$. Let $\theta_0 \in \Theta$ be the true value of the parameter, i.e. $P = P_{\theta_0}$. Suppose that the members $P_\theta$ of $\mathcal{Q}$ all have densities $p_\theta$ dominated by a measure $\mu$, and that

$$\dot{\ell}_{\theta_0} \equiv \frac{\partial}{\partial \theta} \log p_\theta \Big|_{\theta=\theta_0},$$

where $\dot{\ell}_{\theta_0} \in L_2^0(P)$, $P\|\dot{\ell}_\theta - \dot{\ell}_{\theta_0}\|^2 \to 0$ as $\theta \to \theta_0$, and the meaning of the extension of $L_2^0(P)$ to vectors of random variables is obvious. As discussed in the paragraphs following (3.1), the tangent set $\dot{\mathcal{Q}}_P \equiv \{h'\dot{\ell}_{\theta_0} : h \in \mathbb{R}^k\}$ contains all the information in the score $\dot{\ell}_{\theta_0}$, and, moreover, it is not hard to verify that $\dot{\mathcal{Q}}_P$ represents $\mathcal{Q}$.

Thus one-dimensional submodels are sufficient to represent all finite-dimensional submodels. Moreover, since semiparametric efficiency is assessed by examining the information for the worst finite-dimensional submodel, one-dimensional submodels are sufficient for semiparametric models in general, including models with infinite-dimensional parameters.

Now if $\{P_t : t \in N_\epsilon\}$ and $g \in \dot{\mathcal{P}}_P$ satisfy (3.1), then for any $a \geq 0$, everything will also hold when $\epsilon$ is replaced by $\epsilon/a$ and $g$ is replaced by $ag$. Thus we can usually assume, without a significant loss in generality, that a tangent set $\dot{\mathcal{P}}_P$ is a *cone*, i.e., a set that is closed under multiplication by nonnegative scalars. We will also frequently find it useful to replace a tangent set with its closed linear span, or to simply assume that the tangent set is closed under limits of linear combinations, in which case it becomes a tangent space.

We now review and generalize several additional important concepts from Section 3.1. For an arbitrary model parameter $\psi : \mathcal{P} \mapsto \mathbb{D}$, consider the fairly general setting where $\mathbb{D}$ is a Banach space $\mathbb{B}$. In this case, $\psi$ is *differentiable at $P$ relative to the tangent set $\dot{\mathcal{P}}_P$* if, for every smooth one-dimensional submodel $\{P_t\}$ with tangent $g \in \dot{\mathcal{P}}_P$, $d\psi(P_t)/(dt)|_{t=0} = \dot{\psi}_P(g)$ for some bounded linear operator $\dot{\psi}_P : \dot{\mathcal{P}}_P \mapsto \mathbb{B}$.

When $\dot{\mathcal{P}}_P$ is a linear space, it is a subspace of the Hilbert space $L_2^0(P)$ and some additional results follow. To begin with, the Riesz representation theorem yields that for every $b^* \in \mathbb{B}^*$, $b^*\dot{\psi}_P(g) = P\left[\tilde{\psi}_P(b^*)g\right]$ for some operator $\tilde{\psi}_P : \mathbb{B}^* \mapsto \overline{\text{lin}}\,\dot{\mathcal{P}}_P$. Note also that for any $g \in \dot{\mathcal{P}}_P$ and $b^* \in \mathbb{B}^*$, we also have $b^*\dot{\psi}_P(g) = \langle g, \dot{\psi}_P^*(b^*)\rangle$, where $\langle\cdot,\cdot\rangle$ is the inner product on $L_2^0(P)$ and $\dot{\psi}_P^*$ is the adjoint of $\dot{\psi}_P$. Thus the operator $\tilde{\psi}_P$ is precisely $\dot{\psi}_P^*$. In this case, $\tilde{\psi}_P$ is the *efficient influence function*. The justification for the qualifier "efficient" will be given in Theorems 18.3 and 18.4 in the next section.

An estimator sequence $\{T_n\}$ for a parameter $\psi(P)$ is *asymptotically linear* if there exists an *influence function* $\check{\psi}_P : \mathcal{X} \mapsto \mathbb{B}$ such that $\sqrt{n}(T_n - \psi(P)) - \sqrt{n}\mathbb{P}_n\check{\psi}_P \xrightarrow{\text{P}} 0$. The estimator $T_n$ is *regular* at $P$ relative to $\dot{\mathcal{P}}_P$ if for every smooth one-dimensional submodel $\{P_t\} \subset \mathcal{P}$ and every sequence $t_n$ with $t_n = O(n^{-1/2})$, $\sqrt{n}(T_n - \psi(P_{t_n})) \overset{P_n}{\rightsquigarrow} Z$, for some tight Borel random element $Z$, where $P_n \equiv P_{t_n}$.

There are a number of ways to establish regularity, but when $\mathbb{B} = \ell^\infty(\mathcal{H})$, for some set $\mathcal{H}$, and $T_n$ is asymptotically linear, the conditions for regularity can be expressed as properties of the influence function, as verified in the theorem below. Note that by Proposition 18.14 in Section 18.4 below, we only need to consider the influence function as evaluated for the subset of linear functionals $\mathbb{B}' \subset \mathbb{B}^*$ that are coordinate projections. These projections are defined by the relation $\dot{\psi}_P(g)(h) = P[\tilde{\psi}_P(h)g]$ for all $h \in \mathcal{H}$.

THEOREM 18.1 *Assume $T_n$ and $\psi(P)$ are in $\ell^\infty(\mathcal{H})$, that $\psi$ is differentiable at $P$ relative to the tangent space $\dot{\mathcal{P}}_P$ with efficient influence function $\tilde{\psi}_P : \mathcal{H} \mapsto L_2^0(P)$, and that $T_n$ is asymptotically linear for $\psi(P)$, with influence function $\check{\psi}_P$. For each $h \in \mathcal{H}$, let $\check{\psi}_P^\bullet(h)$ be the projection of $\check{\psi}_P(h)$ onto $\dot{\mathcal{P}}_P$. Then the following are equivalent:*

*(i) The class $\mathcal{F} \equiv \{\check{\psi}_P(h) : h \in \mathcal{H}\}$ is $P$-Donsker and $\check{\psi}_P^\bullet(h) = \tilde{\psi}_P(h)$ almost surely for all $h \in \mathcal{H}$.*

*(ii) $T_n$ is regular at $P$.*

Before giving the proof, we highlight an important rule for verifying that a candidate influence function $\check{\psi}_P : \mathcal{H} \mapsto L_2^0(P)$ is an efficient influence function. This rule can be extracted from preceding arguments and is summarized in the following proposition, the proof of which is saved as an exercise (see Exercise 18.5.1):

PROPOSITION 18.2 *Assume $\psi : \mathcal{P} \mapsto \ell^\infty(\mathcal{H})$ is differentiable at $P$ relative to the linear tangent set $\dot{\mathcal{P}}_P$, with bounded linear derivative $\dot{\psi}_P : \dot{\mathcal{P}}_P \mapsto \ell^\infty(\mathcal{H})$. Then $\check{\psi}_P : \mathcal{H} \mapsto L_2^0(P)$ is an efficient influence function if and only if the following both hold:*

*(i) $\check{\psi}_P(h)$ is in the closed linear span of $\dot{\mathcal{P}}_P$ for all $h \in \mathcal{H}$, and*

*(ii)* $\dot{\psi}_P(g)(h) = P[\check{\psi}_P(h)g]$ *for all* $h \in \mathcal{H}$ *and* $g \in \dot{\mathcal{P}}_P$.

This simple proposition is a primary ingredient in a "calculus" of semiparametric efficient estimators and was mentioned in a less general form in Chapter 3. A second important ingredient is making sure that $\dot{\mathcal{P}}_P$ is rich enough to represent all smooth finite-dimensional submodels of $\mathcal{P}$ (see Exercise 3.5.1 for a discussion of this regarding the Cox model for right-censored data). This issue was also discussed a few paragraphs previously but is sufficiently important to bare repeating.

**Proof of Theorem 18.1.** Suppose $\mathcal{F}$ is $P$-Donsker. Let $\{P_t\}$ be any smooth one-dimensional submodel with tangent $g \in \dot{\mathcal{P}}_P$, and let $t_n$ be any sequence with $\sqrt{n}t_n \to k$, for some finite constant $k$. Then $P_n = P_{t_n}$ satisfies (11.4) for $h = g$, and thus, by Theorem 11.12,

$$\sqrt{n}\mathbb{P}_n\check{\psi}_P(\cdot) \overset{P_n}{\rightsquigarrow} \mathbb{G}\check{\psi}_P(\cdot) + P\left[\check{\psi}_P(\cdot)g\right] = \mathbb{G}\check{\psi}_P(\cdot) + P\left[\check{\psi}_P^{\bullet}(\cdot)g\right]$$

in $\ell^\infty(\mathcal{H})$. The last equality follows from the fact that $g \in \dot{\mathcal{P}}_P$. Let $Y_n = \left\|\sqrt{n}(T_n - \psi(P)) - \sqrt{n}\mathbb{P}_n\check{\psi}_P\right\|_{\mathcal{H}}$, and note that $Y_n \overset{\mathrm{P}}{\to} 0$ by the asymptotic linearity assumption.

At this point, we note that it is really no loss of generality to assume that the measurable sets for $P_n$ and $P^n$ (applied to the data $X_1, \ldots, X_n$) are both the same for all $n \geq 1$. Using this assumption, we now have by Theorem 11.14 that $Y_n \overset{P_n}{\to} 0$. Combining this with the differentiability of $\psi$, we obtain that

$$\begin{aligned} \sqrt{n}(T_n - \psi(P_n))(\cdot) &= \sqrt{n}(T_n - \psi(P))(\cdot) \\ &\quad -\sqrt{n}(\psi(P_n) - \psi(P))(\cdot) \\ &\overset{P_n}{\rightsquigarrow} \mathbb{G}\check{\psi}_P(\cdot) + P\left[\left(\check{\psi}_P^{\bullet}(\cdot) - \tilde{\psi}_P(\cdot)\right)g\right], \end{aligned} \tag{18.2}$$

in $\ell^\infty(\mathcal{H})$.

Suppose now that (ii) holds but we don't know whether $\mathcal{F}$ is $P$-Donsker. The fact that $T_n$ is regular implies that $\sqrt{n}(T_n - \psi(P)) \rightsquigarrow Z$, for some tight process $Z$. The asymptotic linearity of $T_n$ now forces $\sqrt{n}\mathbb{P}_n\check{\psi}_P(\cdot) \rightsquigarrow Z$ also, and this yields that $\mathcal{F}$ is in fact $P$-Donsker. Suppose, however, that for some $h \in \mathcal{H}$ we have $\tilde{g}(h) \equiv \check{\psi}_P^{\bullet}(h) - \tilde{\psi}_P(h) \neq 0$. Since the choice of $g$ in the arguments leading up to (18.2) was arbitrary, we can choose $g = a\tilde{g}$ for any $a > 0$ to yield

$$\sqrt{n}(T_n(h) - \psi(P_n)(h)) \overset{P_n}{\rightsquigarrow} \mathbb{G}\check{\psi}_P(h) + aP\tilde{g}^2. \tag{18.3}$$

Thus we can easily have different limiting distributions by choosing different values of $a$. This means that $T_n$ is not regular. Thus we have proved that (i) holds by contradiction.

Assume now that (i) holds. Using once again the arguments preceding (18.2), we obtain for arbitrary choices of $g \in \dot{\mathcal{P}}_P$ and constants $k$, that

$$\sqrt{n}(T_n - \psi(P_n)) \overset{P_n}{\rightsquigarrow} \mathbb{G}\check{\psi}_P$$

in $\ell^\infty(\mathcal{H})$. Now relax the assumption that $\sqrt{n}t_n \to k$ to $\sqrt{n}t_n = O(1)$ and allow $g$ to arbitrary as before. Under this weaker assumption, we have that for every subsequence $n'$, there exists a further subsequence $n''$ such that $\sqrt{n''}t_{n''} \to k$, for some finite $k$, as $n'' \to \infty$. Arguing along this subsequence, our previous arguments can all be recycled to verify that

$$\sqrt{n''}(T_{n''} - \psi(P_{n''})) \overset{P_{n''}}{\rightsquigarrow} \mathbb{G}\check{\psi}_P$$

in $\ell^\infty(\mathcal{H})$, as $n'' \to \infty$.

Fix $f \in C_b(\ell^\infty(\mathcal{H}))$, recall the portmanteau theorem (Theorem 7.6), and define $Z_n \equiv \sqrt{n}(T_n - \psi(P_n))$. What we have just proven is that every subsequence $n'$ has a further subsequence $n''$ such that $E^* f(Z_{n''}) \to Ef(Z)$, as $n'' \to \infty$, where $Z \equiv \mathbb{G}\check{\psi}_P$. This of course implies that $E^* f(Z_n) \to Ef(Z)$, as $n \to \infty$. Reapplication of the portmanteau theorem now yields $Z_n \overset{P_n}{\rightsquigarrow} Z$. Thus $T_n$ is regular, and the proof is complete for both directions.□

We note that in the context of Theorem 18.1 above, a non-regular estimator $T_n$ has a serious defect. From (18.3), we see that for any $M < \infty$ and $\epsilon > 0$, there exists a smooth one-dimensional submodel $\{P_t\}$ such that

$$\mathrm{P}\left(\left\|\sqrt{n}(T_n - \psi(P_n))\right\|_{\mathcal{H}} > M\right) > 1 - \epsilon.$$

Thus the estimator $T_n$ has arbitrarily poor performance for certain submodels which are represented by $\dot{\mathcal{P}}_P$. Hence regularity is not just a mathematically convenient definition, but it reflects, even in infinite-dimensional settings, a certain intuitive reasonableness about $T_n$. This does not mean that nonregular estimators are never useful, because they can be, especially when the parameter $\psi(P)$ is not $\sqrt{n}$-consistent. Nevertheless, regular estimators are very appealing when they are available.

## 18.2 Efficiency

We now turn our attention to the question of efficiency in estimating general Banach-valued parameters. We first present general optimality results and then characterize efficient estimators in the special Banach space $\ell^\infty(\mathcal{H})$. We then consider efficiency of Hadamard-differentiable functionals of efficient parameters and show how to establish efficiency of estimators in $\ell^\infty(\mathcal{H})$ from efficiency of all one-dimensional components. We also consider the related issue of efficiency in product spaces.

The following two theorems, which characterize optimality in Banach spaces, are the key results of this section. For a Borel random element $Y$, let $L(Y)$ denote the law of $Y$ (as in Section 7.1), and let $*$ denote the convolution operation. Also define a function $u : \mathbb{B} \mapsto [0, \infty)$ to be *subconvex*

if, for every $b \in \mathbb{B}$, $u(b) \geq 0 = u(0)$ and $u(b) = u(-b)$, and also, for every $c \in \mathbb{R}$, the set $\{b \in \mathbb{B} : u(b) \leq c\}$ is convex and closed. A simple example of a subconvex function is the norm $\|\cdot\|$ for $\mathbb{B}$. Here are the theorems:

THEOREM 18.3 *(Convolution theorem) Assume that $\psi : \mathcal{P} \mapsto \mathbb{B}$ is differentiable at $P$ relative to the tangent space $\dot{\mathcal{P}}_P$, with efficient influence function $\tilde{\psi}_P$. Assume that $T_n$ is regular at $P$ relative to $\dot{\mathcal{P}}_P$, with $Z$ being the tight weak limit of $\sqrt{n}(T_n - \psi(P))$ under $P$. Then $L(Z) = L(Z_0) * L(M)$, where $M$ is some Borel random element in $\mathbb{B}$, and $Z_0$ is a tight Gaussian process in $\mathbb{B}$ with covariance $P[(b_1^* Z_0)(b_2^* Z_0)] = P\left[\tilde{\psi}_P(b_1^*)\tilde{\psi}_P(b_2^*)\right]$ for all $b_1^*, b_2^* \in \mathbb{B}^*$.*

THEOREM 18.4 *Assume the conditions of Theorem 18.3 hold and that $u : \mathbb{B} \mapsto [0, \infty)$ is subconvex. Then*

$$\limsup_{n \to \infty} \mathrm{E}_* u\left(\sqrt{n}(T_n - \psi(P))\right) \geq \mathrm{E}u(Z_0),$$

*where $Z_0$ is as defined in Theorem 18.3.*

We omit the proofs, but they can be found in Section 5.2 of BKRW.

The previous two theorems characterize optimality of regular estimators in terms of the limiting process $Z_0$, which is a tight, mean zero Gaussian process with covariance obtained from the efficient influence function. This can be viewed as an asymptotic generalization of the Cramér-Rao lower bound. We say that an estimator $T_n$ is *efficient* if it is regular and the limiting distribution of $\sqrt{n}(T_n - \psi(P))$ is $Z_0$, i.e., $T_n$ achieves the optimal lower bound. The following proposition, the proof of which is given in Section 4, assures us that $Z_0$ is fully characterized by the distributions of $b^* Z_0$ for $b^*$ ranging over all of $\mathbb{B}^*$:

PROPOSITION 18.5 *Let $X_n$ be an asymptotically tight sequence in a Banach space $\mathbb{B}$ and assume $b^* X_n \rightsquigarrow b^* X$ for every $b^* \in \mathbb{B}^*$ and some tight, Gaussian process $X$ in $\mathbb{B}$. Then $X_n \rightsquigarrow X$.*

The next result assures us that Hadamard differentiable functions of efficient estimators are also asymptotically efficient:

THEOREM 18.6 *Assume that $\psi : \mathcal{P} \mapsto \mathbb{B}$ is differentiable at $P$ relative to the tangent space $\dot{\mathcal{P}}_P$—with derivative $\dot{\psi}_P g$, for every $g \in \dot{\mathcal{P}}_P$, and efficient influence function $\tilde{\psi}_P$—and takes its values in a subset $\mathbb{B}_\phi$. Suppose also that $\phi : \mathbb{B}_\phi \subset \mathbb{B} \mapsto \mathbb{E}$ is Hadamard differentiable at $\psi(P)$ tangentially to $\mathbb{B}_0 \equiv \overline{lin}\, \dot{\psi}_P(\dot{\mathcal{P}}_P)$. Then $\phi \circ \psi : \mathcal{P} \mapsto \mathbb{E}$ is also differentiable at $P$ relative to $\dot{\mathcal{P}}_P$. If $T_n$ is a sequence of estimators with values in $\mathbb{B}_\phi$ that is efficient at $P$ for estimating $\psi(P)$, then $\phi(T_n)$ is efficient at $P$ for estimating $\phi \circ \psi(P)$.*

**Proof.** Let $\phi'_{\psi(P)} : \mathbb{B} \mapsto \mathbb{E}$ be the derivative of $\phi$. Note that for any $g \in \dot{\mathcal{P}}_P$ and any submodel $\{P_t\}$ with tangent $g$, we have by the specified Hadamard differentiability of $\phi$ that

$$\frac{\phi \circ \psi(P_t) - \phi \circ \psi(P)}{t} \quad \to \quad \phi'_{\psi(P)} \dot{\psi}_P g, \tag{18.4}$$

as $t \to 0$. Thus $\phi \circ \psi : \mathcal{P} \mapsto \mathbb{E}$ is differentiable at $P$ relative to $\dot{\mathcal{P}}_P$.

For any chosen submodel $\{P_t\}$ with tangent $g \in \dot{\mathcal{P}}_P$, define $P_n \equiv P_{1/\sqrt{n}}$. By the efficiency of $T_n$, we have that $\sqrt{n}(T_n - \psi(P_n)) \overset{P_n}{\rightsquigarrow} Z_0$, where $Z_0$ has the optimal, mean zero, tight Gaussian limiting distribution. By the delta method (Theorem 2.8), we now have that $\sqrt{n}(\phi(T_n) - \phi \circ \psi(P_n)) \overset{P_n}{\rightsquigarrow} \phi'_{\psi(P)} Z_0$. Since the choice of $\{P_t\}$ was arbitrary, we now know that $\phi(T_n)$ is regular and also that $\sqrt{n}(\phi(T_n) - \phi \circ \psi(P)) \rightsquigarrow \phi'_{\psi(P)} Z_0$. By Theorem 18.3, $P[(e_1^* \phi'_{\psi(P)} Z_0)(e_2^* \phi'_{\psi(P)} Z_0)] = P[\tilde{\psi}_P(e_1^* \phi'_{\psi(P)}) \tilde{\psi}_P(e_2^* \phi'_{\psi(P)})]$, for all $e_1^*, e_2^* \in \mathbb{E}^*$. Thus the desired result now follows from (18.4) and the definition of $\tilde{\psi}_P$ since $P[\tilde{\psi}_P(e^* \phi'_{\psi(P)}) g] = e^* \phi'_{\psi(P)} \dot{\psi}_P(g)$ for every $e^* \in \mathbb{E}^*$ and $g \in \overline{\text{lin}}\, \dot{\mathcal{P}}_P$.□

The following very useful theorem completely characterizes efficient estimators of Euclidean parameters. The theorem is actually Lemma 25.23 of van der Vaart (1998), and the proof, which we omit, can be found therein:

THEOREM 18.7 *Let $T_n$ be an estimator for a parameter $\psi : \mathcal{P} \mapsto \mathbb{R}^k$, where $\psi$ is differentiable at $P$ relative to the tangent space $\dot{\mathcal{P}}_P$ with $k$-variate efficient influence function $\tilde{\psi}_P \in L_2^0(P)$. Then the following are equivalent:*

*(i) $T_n$ is efficient at $P$ relative to $\dot{\mathcal{P}}_P$, and thus the limiting distribution of $\sqrt{n}(T_n - \psi(P))$ is mean zero normal with covariance $P[\tilde{\psi}_P \tilde{\psi}'_P]$.*

*(ii) $T_n$ is asymptotically linear with influence function $\tilde{\psi}_P$.*

The next theorem we present endeavors to extend the above characterization of efficient estimators to more general parameter spaces of the form $\ell^\infty(\mathcal{H})$:

THEOREM 18.8 *Let $T_n$ be an estimator for a parameter $\psi : \mathcal{P} \mapsto \ell^\infty(\mathcal{H})$, where $\psi$ is differentiable at $P$ relative to the tangent space $\dot{\mathcal{P}}_P$ with efficient influence function $\tilde{\psi}_P : \mathcal{H} \mapsto L_2^0(P)$. Let $\mathcal{F} \equiv \{\tilde{\psi}_P(h) : h \in \mathcal{H}\}$. Then the following are equivalent:*

*(a) $T_n$ is efficient at $P$ relative to $\dot{\mathcal{P}}_P$ and at least one of the following holds:*

  *(i) $T_n$ is asymptotically linear.*

  *(ii) $\mathcal{F}$ is $P$-Donsker for some version of $\tilde{\psi}_P$.*

*(b) For some version of $\tilde{\psi}_P$, $T_n$ is asymptotically linear with influence function $\tilde{\psi}_P$ and $\mathcal{F}$ is $P$-Donsker.*

*(c) $T_n$ is regular and asymptotically linear with influence function $\check{\psi}_P$ such that $\{\check{\psi}_P(h) : h \in \mathcal{H}\}$ is $P$-Donsker and $\check{\psi}_P(h) \in \dot{\mathcal{P}}_P$ for all $h \in \mathcal{H}$.*

We are assuming tacitly that $\{\tilde{\psi}_P(h) : h \in \mathcal{H}\}$ is a stochastic process. Two stochastic processes $X$ and $\tilde{X}$ are versions of each other when $X(h) = \tilde{X}(h)$ almost surely for every $h \in \mathcal{H}$. Note that this is a slightly different definition than the one used in Section 7.1 for Borel random variables, but the concepts are equivalent when the processes are both Borel measurable and live on $\ell^\infty(\mathcal{H})$. This equivalence follows since finite-dimensional distributions determine the full distribution for tight stochastic processes. However, in the generality of the above theorem, we need to be careful since $\tilde{\psi}_P$ may not be Borel measurable when $\mathcal{H}$ is infinite and different versions may have different properties.

The theorem gives us several properties of efficient estimators that can be useful for a number of things, including establishing efficiency. In particular, conclusion (c) tells us that a simple method for establishing efficiency of $T_n$ requires only that $T_n$ be regular and asymptotically linear with an influence function that is contained in a Donsker class for which the individual components $\check{\psi}_P(h)$ are contained in the tangent space for all $h \in \mathcal{H}$. The theorem also tells us that if $T_n$ is efficient, only one of (i) or (ii) in (a) is required and the other will follow. This means, for example, that if $T_n$ is efficient and $\mathcal{F}$ is not $P$-Donsker for any version of $\tilde{\psi}_P$, then $T_n$ must not be asymptotically linear. Also note that the requirement that $\mathcal{F}$ is $P$-Donsker collapses to requiring that $\|\tilde{\psi}_P\|_{P,2} < \infty$ when $\mathcal{H}$ is finite and we are therefore in the setting of Theorem 18.7. However, such a requirement is not needed in the statement of Theorem 18.7 since $\|\tilde{\psi}_P\|_{P,2} < \infty$ automatically follows from the required differentiability of $\psi$ when $\psi \in \mathbb{R}^k$. This follows since the Riesz representation theorem assures us that $\tilde{\psi}_P$ is in the closed linear span of $\dot{\mathcal{P}}_P$ which is a subset of $L_2(P)$.

**Proof of Theorem 18.8.** Assume (b), and note that Theorem 18.1 now implies that $T_n$ is regular. Efficiency now follows immediately from Theorems 18.3 and 18.4 and the definition of efficiency. Thus (a) follows.

Now assume (a) and (i). Since the definition of efficiency includes regularity, we know by Theorem 18.1 that $T_n$ is asymptotically linear with influence function $\check{\psi}_P$ having projection $\tilde{\psi}_P$ on $\dot{\mathcal{P}}_P$. By Theorem 18.3, $\check{\psi}_P(h) - \tilde{\psi}_P(h) = 0$ almost surely, for all $h \in \mathcal{H}$. Since $\{\check{\psi}_P(h) : h \in \mathcal{H}\}$ is $P$-Donsker by the asymptotic linearity and efficiency of $T_n$, we now have that $\check{\psi}_P$ is a version of $\tilde{\psi}_P$ for which $\mathcal{F}$ is $P$-Donsker. This yields (b). Note that in applying Theorem 18.3, we only needed to consider the evaluation functionals in $\mathbb{B}^*$ (i.e., $b_h^*(b) \equiv b(h)$, where $h \in \mathcal{H}$) and not all of $\mathbb{B}^*$, since a tight, mean zero Gaussian process on $\ell^\infty(\mathcal{H})$ is completely determined by its covariance function $(h_1, h_2) \mapsto P[Z_0(h_1)Z_0(h_2)]$ (see Proposition 18.14 below).

Now assume (a) and (ii). The regularity of $T_n$ and the fact that $\mathcal{F}$ is $P$-Donsker yields that both $\sqrt{n}(T_n - \psi(P)$ and $\sqrt{n}\mathbb{P}_n\tilde{\psi}_P$ are asymptotically tight. Thus, for verifying the desired asymptotic linearity, it is sufficient to verify it for all finite-dimensional subsets of $\mathcal{H}$, i.e., we need to verify that

$$\sup_{h\in\mathcal{H}_0}\left|\sqrt{n}(T_n(h)-\psi(P)(h))-\sqrt{n}\mathbb{P}_n\tilde{\psi}_P(h)\right| \quad = \quad o_P(1), \tag{18.5}$$

for all finite subsets $\mathcal{H}_0 \subset \mathcal{H}$. This holds by Theorem 18.7, and thus the desired conclusions follow.

Note that (c) trivially follows from (b). Moreover, (b) follows from (c) by Theorem 18.1 and the regularity of $T_n$. This completes the proof.□

The following theorem tells us that pointwise efficiency implies uniform efficiency under weak convergence. This fairly deep result is quite useful in applications.

THEOREM 18.9 *Let $T_n$ be an estimator for a parameter $\psi : \mathcal{P} \mapsto \ell^\infty(\mathcal{H})$, where $\psi$ is differentiable at $P$ relative to the tangent space $\dot{\mathcal{P}}_P$ with efficient influence function $\tilde{\psi}_P : \mathcal{H} \mapsto L_2^0(P)$. The following are equivalent:*

*(a) $T_n$ is efficient for $\psi(P)$.*

*(b) $T_n(h)$ is efficient for $\psi(P)(h)$, for every $h \in \mathcal{H}$, and $\sqrt{n}(T_n - \psi(P))$ is asymptotically tight under $P$.*

The proof of this theorem makes use of the following deep lemma, the proof of which is given in Section 4 below.

LEMMA 18.10 *Suppose that $\psi : \mathcal{P} \mapsto \mathbb{D}$ is differentiable at $P$ relative to the tangent space $\dot{\mathcal{P}}_P$ and that $d'T_n$ is asymptotically efficient at $P$ for estimating $d'\psi(P)$ for every $d'$ in a subset $\mathbb{D}' \subset \mathbb{D}^*$ which satisfies*

$$\|d\| \quad \leq \quad c \sup_{d'\in\mathbb{D}',\, \|d'\|\leq 1} |d'(d)|, \tag{18.6}$$

*for some constant $c < \infty$. Then $T_n$ is asymptotically efficient at $P$ provided $\sqrt{n}(T_n - \psi(P))$ is asymptotically tight under $P$.*

**Proof of Theorem 18.9.** That (a) implies (b) is obvious. Now assume (b), and let $\mathbb{D} = \ell^\infty(\mathcal{H})$ and $\mathbb{D}'$ be the set of all coordinate projections $d \mapsto d_h^* d \equiv d(h)$ for every $h \in \mathcal{H}$. Since the uniform norm on $\ell^\infty(\mathcal{H})$ is trivially equal to $\sup_{d'\in\mathbb{D}'} |d'd|$ and all $d' \in \mathbb{D}'$ satisfy $\|d'\| = 1$, Condition (18.6) is easily satisfied. Since $\sqrt{n}(T_n - \psi(P))$ is asymptotically tight by assumption, all of the conditions of Lemma 18.10 are satisfied. Hence $T_n$ is efficient, and the desired conclusions follow.□

We close this section with an interesting corollary of Lemma 18.10 that provides a remarkably simple connection between marginal and joint efficiency on product spaces:

THEOREM 18.11 *Suppose that $\psi_j : \mathcal{P} \mapsto \mathbb{D}_j$ is differentiable at $P$ relative to the tangent space $\dot{\mathcal{P}}_P$, and suppose that $T_{n,j}$ is asymptotically efficient at $P$ for estimating $\psi_j(P)$, for $j = 1, 2$. Then $(T_{n,1}, T_{n,2})$ is asymptotically efficient at $P$ for estimating $(\psi_1(P), \psi_2(P))$.*

Thus marginal efficiency implies joint efficiency even though marginal weak convergence does not imply joint weak convergence! This is not quite so surprising as it may appear at first. Consider the finite-dimensional setting where $\psi_j(P) \in \mathbb{R}$ for $j = 1, 2$. If $T_{n,j}$ is efficient for $\psi_j(P)$, for each $j = 1, 2$, then Theorem 18.7 tells us that $(T_{n,1}, T_{n,2})$ is asymptotically linear with influence function $(\tilde{\psi}_{1,P}, \tilde{\psi}_{2,P})$. Thus the limiting joint distribution will in fact be the optimal bivariate Gaussian distribution. The above theorem can be viewed as an infinite-dimension generalization of this finite-dimensional phenomenon.

**Proof of Theorem 18.11.** Let $\mathbb{D}'$ be the set of all maps $(d_1, d_2) \mapsto d_j^* d_j$ for $d_j^* \in \mathbb{D}_j$ and $j$ equal to either 1 or 2. Note that by the Hahn-Banach theorem (see Corollary 6.7 of Conway, 1990), $\|d_j\| = \sup\{|d_j^* d_j| : \|d_j^*\| = 1, d_j^* \in \mathbb{D}_j^*\}$. Thus the product norm $\|(d_1, d_2)\| = \|d_1\| \vee \|d_2\|$ satisfies Condition (18.6) of Lemma 18.10 with $c = 1$. Hence the desired conclusion follows.□

## 18.3 Optimality of Tests

In this section, we study testing of the null hypothesis

$$H_0 : \psi(P) \leq 0 \tag{18.7}$$

versus the alternative $H_1 : \psi(P) > 0$ for a one-dimensional parameter $\psi(P)$. The basic conclusion we will endeavor to show is that a test based on an asymptotically optimal estimator for $\psi(P)$ will, in a meaningful way, be asymptotically optimal. Note that null hypotheses of the form $H_{01} : \psi(P) \leq \psi_0$ can trivially be rewritten in the form given in (18.7) by replacing $P \mapsto \psi(P)$ with $P \mapsto \psi(P) - \psi_0$. For dimensions higher than one, coming up with a satisfactory criteria forming up with a satisfactory criteria for optimality of tests is difficult (see the discussion in Section 15.2 of van der Vaart, 1998) and we will not pursue the higher dimensional setting in this book.

For a given model $\mathcal{P}$ and measure $P$ on the boundary of the null hypothesis where $\psi(P) = 0$, we are interested in studying the "local asymptotic power" in a neighborhood of $P$. These neighborhoods are of size $1/\sqrt{n}$ and are the appropriate magnitude when considering sample size computation for $\sqrt{n}$ consistent parameter estimates. Consider for example the univariate normal setting where the data are i.i.d. $N(\mu, \sigma^2)$. A natural choice of test for $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$ is the indicator of whether $T_n = \sqrt{n}\bar{x}/s_n > z_{1-\alpha}$, where $\bar{x}$ and $s_n$ are the sample mean and standard deviation from an i.i.d. sample $X_1, \ldots, X_n$, $z_q$ is the $q$th quantile of a standard normal, and $\alpha$ is the chosen size of this one-sided test. For any $\mu > 0$, $T_n$ diverges to infinity with probability 1. However, if $\mu = k/\sqrt{n}$ for some finite $k$, then $T_n \rightsquigarrow N(k, 1)$. Thus we can derive non-trivial power functions

only for shrinking "contiguous alternatives" in a $1/\sqrt{n}$ neighborhood of zero. In this case, since $\tilde{\psi}_P = X$ and the corresponding one-dimensional submodel $\{P_t\}$ must satisfy $\partial\psi(P_t)/(\partial t)|_{t=1} = k$, we know that the score function $g$ corresponding to $\{P_t\}$ must be $g(X) = kX/\sigma$. Hence, in this example, we can easily express the local alternative sequence in terms of the score function rather than $k$.

Thus it makes sense in general to study the performance of tests under contiguous alternatives defined by one-dimensional submodels corresponding to score functions. Accordingly, for a given element $g$ of a tangent set $\dot{\mathcal{P}}_P$, let $t \mapsto P_{t,g}$ be a one-dimensional submodel which is differentiable in quadratic mean at $P$ with score function $g$ along which $\psi$ is differentiable, i.e.,

$$\frac{\psi(P_{t,g}) - \psi(P)}{t} \to P[\tilde{\psi}_P g],$$

as $t \downarrow 0$. For each such $g$ for which $P[\tilde{\psi}_P g] > 0$, we can see that when $\psi(P) = 0$, the submodel $\{P_{t,g}\}$ belongs to $H_1 : \psi(P) > 0$ for all sufficiently small $t > 0$. We will therefore consider power over contiguous alternatives of the form $\{P_{h/\sqrt{n},g}\}$ for $h > 0$.

Before continuing, we need to define a *power function*. For a subset $\mathcal{Q} \subset \mathcal{P}$ containing $P$, a power function $\pi : \mathcal{Q} \mapsto [0,1]$ at level $\alpha$ is a function on probability measures that satisfies $\pi(Q) \leq \alpha$ for all $Q \in \mathcal{Q}$ for which $\psi(Q) \leq 0$. We say that a sequence of power function $\{\pi_n\}$ has asymptotic level $\alpha$ if $\limsup_{n\to\infty} \pi_n(Q) \leq \alpha$ for every $Q \in \mathcal{Q} : \psi(Q) \leq 0$. The power function for a level $\alpha$ hypothesis test of $H_0$ is the probability of rejecting $H_0$ under $Q$. Hence statements about power functions can be viewed as statements about hypothesis tests. Here is our main result:

THEOREM 18.12 *Let $\psi : \mathcal{P} \mapsto \mathbb{R}$ be differentiable at $P$ relative to the tangent space $\dot{\mathcal{P}}_P$ with efficient influence function $\tilde{\psi}_P$, and suppose $\psi(P) = 0$. Then, for every sequence of power functions $P \mapsto \pi_n(P)$ of asymptotic level $\alpha$ tests for $H_0 : \psi(P) \leq 0$, and for every $g \in \dot{\mathcal{P}}_P$ with $P[\tilde{\psi}_P g] > 0$ and every $h > 0$,*

$$\limsup_{n\to\infty} \pi_n(P_{h/\sqrt{n},g}) \leq 1 - \Phi\left[z_{1-\alpha} - h\frac{P[\tilde{\psi}_P g]}{\sqrt{P[\tilde{\psi}_P^2]}}\right].$$

This is a minor modification of Theorem 25.44 in Section 25.6 of van der Vaart (1998) and the proof is given therein. While there are some differences in notation, the substantive modification is that van der Vaart requires the power functions to have level $\alpha$ for each $n$, whereas our version only require the levels to be asymptotically $\alpha$. This does not affect the proof, and we omit the details. An advantage of this modification is that it permits the use of approximate hypothesis tests, such as those which depend on the central limit theorem, whose level for fixed $n$ may not be exactly $\alpha$ but whose asymptotic level is known to be $\alpha$.

An important consequence of Theorem 18.12 is that a test based on an efficient estimator $T_n$ of $\psi(P)$ can achieve the given optimality. To see this, let $S_n^2$ be a consistent estimator of the limiting variance of $\sqrt{n}(T_n - \psi(P))$, and let $\pi_n(Q)$ be the power function defined as the probability that $\sqrt{n}T_n/S_n > z_{1-\alpha}$ under the model $Q$. It is easy to see that this power function has asymptotic power $\alpha$ under the null hypothesis. The following result shows that this procedure is asymptotically optimal:

LEMMA 18.13 *Let $\psi : \mathcal{P} \mapsto \mathbb{R}$ be differentiable at $P$ relative to the tangent space $\dot{\mathcal{P}}_P$ with efficient influence function $\tilde{\psi}_P$, and suppose $\psi(P) = 0$. Suppose the estimator $T_n$ is asymptotically efficient at $P$, and, moreover, that $S_n^2 \xrightarrow{\mathrm{P}} P\tilde{\psi}_P^2$. Then, for every $h \geq 0$ and $g \in \dot{\mathcal{P}}_P$,*

$$\limsup_{n\to\infty} \pi_n(P_{h/\sqrt{n},g}) = 1 - \Phi\left(z_{1-\alpha} - h\frac{P[\tilde{\psi}_P g]}{\sqrt{P\tilde{\psi}_P^2}}\right).$$

**Proof.** By Theorem 18.7 and Part (i) of Theorem 11.14, we have that

$$\frac{\sqrt{n}T_n}{S_n} \overset{P_n}{\rightsquigarrow} Z + h\frac{P[\tilde{\psi}_P g]}{\sqrt{P\tilde{\psi}_P^2}},$$

where $P_n \equiv P_{h/\sqrt{n},g}$ and $Z$ has a standard normal distribution. The desired result now follows.□

Consider, for example, the Mann-Whitney test discussed in Section 12.2.2 for comparing two independent samples of respective sample sizes $m$ and $n$. Let $\mathbb{F}_m$ and $\mathbb{G}_n$ be the respective empirical distributions with corresponding true distributions $F$ and $G$ which we assume to be continuous. The Mann-Whitney statistic is $T_n = \int_{\mathbb{R}} \mathbb{G}_n(s)d\mathbb{F}_m(s) - 1/2$ which is consistent for $\psi(P) = \int_{\mathbb{R}} G(s)dF(s) - 1/2$. We are interested in testing the null hypothesis $H_0 : \psi(P) \leq 0$ versus $H_1 : \psi(P) > 0$.

By Theorem 18.11, we know that $(\mathbb{F}_m, \mathbb{G}_m)$ is jointly efficient for $(F, G)$. Moreover, by Lemma 12.3, we know that $(F, G) \mapsto \int_{\mathbb{R}} G(s)dF(s)$ is Hadamard differentiable. Thus Theorem 18.6 applies, and we obtain that $\int_{\mathbb{R}} \mathbb{G}_n(s)$ $d\mathbb{F}_n(s)$ is asymptotically efficient for $\int_{\mathbb{R}} G(s)dF(s)$. Hence Lemma 18.13 also applies, and we obtain that $T_n$ is optimal for testing $H_0$, provided it is suitably standardized. We know from the discussion in Section 12.2.2 that the asymptotic variance of $\sqrt{n}T_n$ is $1/12$. Thus the test that rejects the null when $\sqrt{12n}T_n$ is greater than $z_{1-\alpha}$ is optimal.

Another simple example is the sign test for symmetry about zero for a sample of real random variables $X_1, \ldots, X_n$ with distribution $F$ that is continuous at zero. The test statistic is $T_n = \int_{\mathbb{R}} \mathrm{sign}(x)d\mathbb{F}_n(x)$, where $\mathbb{F}_n$ is the usual empirical distribution. Using arguments similar to those used in the previous paragraphs, it can be shown that $T_n$ is an asymptotically efficient estimator for $\psi(P) = \int_{\mathbb{R}} \mathrm{sign}(x)dF(x) = \mathrm{P}(X > 0) - \mathrm{P}(X <$

0). Thus, by Lemma 18.13 above, the sign test is asymptotically optimal for testing the null hypothesis $H_0 : \mathrm{P}(X > 0) \leq \mathrm{P}(X < 0)$ versus the alternative $H_1 : \mathrm{P}(X > 0) > \mathrm{P}(X < 0)$.

These examples illustrates the general concept that if the parameter of interest is a smooth functional of the underlying distribution functions, then the estimator obtain by substituting the true distributions with the corresponding empirical distributions will be asymptotically optimal, provided we are not willing to make any parametrically restrictive assumptions about the distributions.

## 18.4 Proofs

**Proof of Proposition 18.5.** Let $\mathbb{B}_1^* \equiv \{b^* \in \mathbb{B}^* : \|b^*\| \leq 1\}$ and $\tilde{\mathbb{B}} \equiv \ell^\infty(\mathbb{B}_1^*)$. Note that $(\mathbb{B}, \|\cdot\|) \subset (\tilde{\mathbb{B}}, \|\cdot\|_{\mathbb{B}_1^*})$ by letting $x(b^*) \equiv b^*x$ for every $b^* \in \mathbb{B}^*$ and all $x \in \mathbb{B}$ and recognizing that $\|x\| = \sup_{b^* \in \mathbb{B}^*} |b^*x| = \|x\|_{\mathbb{B}_1^*}$ by the Hahn-Banach theorem (see Corollary 6.7 of Conway, 1990). Thus weak convergence of $X_n$ in $\tilde{\mathbb{B}}$ will imply weak convergence in $\mathbb{B}$ by Lemma 7.8. Since we already know that $X_n$ is asymptotically tight in $\tilde{\mathbb{B}}$, we are done if we can show that all finite-dimensional distributions of $X_n$ converge. Accordingly, let $b_1^*, \ldots, b_m^* \in \mathbb{B}_1^*$ be arbitrary and note that for any $(\alpha_1, \ldots, \alpha_m) \in \mathbb{R}^m$, $\sum_{j=1}^m \alpha_j X_n(b_j^*) = \tilde{b}^* X_n$ for $\tilde{b}^* \equiv \sum_{j=1}^m \alpha_j b_j^* \in \mathbb{B}^*$. Since we know that $\tilde{b}^* X_n \rightsquigarrow \tilde{b}^* X$, we now have that $\sum_{j=1}^m \alpha_j b_j^* X_n \rightsquigarrow \sum_{j=1}^m \alpha_j b_j^* X$. Thus $(X_n(b_1^*), \ldots, X_n(b_m^*))^T \rightsquigarrow (X(b_1^*), \ldots, X(b_m^*))^T$ since $(\alpha_1, \ldots, \alpha_j) \in \mathbb{R}^m$ was arbitrary and $X$ is Gaussian. Since $b_1^*, \ldots, b_m^*$ and $m$ were also arbitrary, we have the result that all finite-dimensional distributions of $X_n$ converge, and the desired conclusion now follows.□

**Proof of Lemma 18.10.** By Part (ii) of Theorem 11.14, we know that $\sqrt{n}(T_n - \psi(P))$ is asymptotically tight under $P_n \equiv P_{1/\sqrt{n}}$ for every differentiable submodel $\{P_t\}$. By the differentiability of $\psi$, we know that $\sqrt{n}(T_n - \psi(P_n))$ is also asymptotically tight under $P_n$. By assumption, we also have that $d'\sqrt{n}(T_n - \psi(P_n))$ is asymptotically linear with covariance $P\tilde{\psi}_P^2(d')$ for every $d' \in \mathbb{D}'$. Since the same result will hold for all finite linear combinations of elements of $\mathbb{D}'$, and since increasing the size of $\mathbb{D}'$ will not negate Condition (18.6), we can replace without loss of generality $\mathbb{D}'$ with $\operatorname{lin}\mathbb{D}'$. By using a minor variation of the proof of Proposition 18.5, we now have that $\sqrt{n}(T_n - \psi(P_n)) \overset{P_n}{\rightsquigarrow} Z$ in $\ell^\infty(\mathbb{D}_1')$, where $\mathbb{D}_1' \equiv \{d^* \in \mathbb{D}' : \|d^*\| \leq 1\}$ and $Z$ is a tight Gaussian process with covariance $P[\tilde{\psi}_P(d_1^*)\tilde{\psi}_P(d_2^*)]$ for all $d_1^*, d_2^* \in \mathbb{D}_1'$. Note that this covariance uniquely defines the distribution of $Z$ as a tight, Gaussian element in $\ell^\infty(\mathbb{D}_1')$ by the linearity of $\mathbb{D}'$ and Lemma 7.3.

Note that by Condition (18.6), we have for every $d \in \mathbb{D}$ that

$$\|d\| \leq c \sup_{d' \in \mathbb{D}', \|d'\| \leq 1} |d'(d)| \leq c\|d\|.$$

We thus have by Lemma 7.8 that the convergence of $\sqrt{n}(T_n - \psi(P_n))$ to $Z$ under $P_n$ also occurs in $\mathbb{D}$. Since Condition 18.6 also implies that $\ell^\infty(\mathbb{D}_1')$ contains $\mathbb{D}$, we have that $Z$ viewed as an element in $\mathbb{D}$ is completely characterized by the covariance $P[\tilde{\psi}_P(d_1^*)\tilde{\psi}_P(d_2^*)]$ for $d_1^*$ and $d_2^*$ ranging over $\mathbb{D}_1'$. This is verified in Proposition 18.14 below. Hence by the assumed differentiability of $\psi$, the covariance of $Z$ also satisfies $P[(d_1^* Z_0)(d_2^* Z_0)] = P[\tilde{\psi}_P(d_1^*)\tilde{\psi}_P(d_2^*)]$ for all $d_1^*, d_2^* \in \mathbb{D}^*$. Hence $T_n$ is regular with $\sqrt{n}(T_n - \psi(P)) \rightsquigarrow Z = Z_0$, where $Z_0$ is the optimal limiting Gaussian process defined in Theorem 18.3. Thus $T_n$ is efficient.□

PROPOSITION 18.14 *Let $X$ be a tight, Borel measurable element in a Banach space $\mathbb{D}$, and suppose there exists a set $\mathbb{D}' \subset \mathbb{D}^*$ such that*

*(i) (18.6) holds for all $d \in \mathbb{D}$ and some constant $0 < c < \infty$ and*

*(ii) $d' \mapsto d'X$ is a Gaussian process on $\ell^\infty(\mathbb{D}_1')$, where $\mathbb{D}_1' \equiv \{d' \in \mathbb{D}' : \|d'\| \le 1\}$.*

*Then $X$ is uniquely defined in law on $\mathbb{D}$.*

Proof. Let $\phi : \mathbb{D} \mapsto \ell^\infty(\mathbb{D}_1')$ be defined by $\phi(d) \equiv \{d'd : d' \in \mathbb{D}_1'\}$, and note that by Exercise 18.5.4 below, both $\phi$ and $\phi^{-1} : \mathbb{D} \mapsto \mathbb{D}$, where

$$\tilde{\mathbb{D}} \equiv \{x \in \ell^\infty(\mathbb{D}_1') : x = \{d'd : d' \in \mathcal{D}_1'\} \text{ for some } d \in \mathbb{D}\},$$

are continuous and linear. Note also that $\tilde{\mathbb{D}}$ is a closed and linear subspace of $\ell^\infty(\mathbb{D}_1')$.

Thus $Y = \phi(X)$ is tight. Now let $d_1', \ldots, d_m' \in \mathbb{D}_1'$ and $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$ be arbitrary. Then by Condition (ii), $\sum_{j=1}^m \alpha_j d_j' X$ is Guassian with variance $\sum_{j,k} \alpha_j \alpha_k P[(d_j' X)(d_k' X)]$. Since the choice of the $\alpha_j$s was arbitrary, $(d_1' X, \ldots, d_m' X)^T$ is multivariate normal with covariance $P[(d_j' X)(d_k' X)]$. Since the choice of $d_1', \ldots, d_m'$ was also arbitrary, we have that all finite-dimensional distributions of $Y$ are multivariate normal. Now the equivalence between (i) and (ii) in Proposition 7.5 implies that $d^* X = d^* \phi^{-1}(Y)$ is Gaussian for all $d^* \in \mathbb{D}^*$. This follows since $\mathrm{P}(X \in \phi^{-1}(\tilde{\mathbb{D}})) = 1$ and, for any $d^* \in \mathbb{D}^*$, $e^* = d^* \phi^{-1}$ is a continuous linear functional on $\tilde{\mathbb{D}}$ that can be extended to a continuous linear functional $\tilde{e}^*$ on all of $\ell^\infty(\mathbb{D}_1')$ with $\tilde{e}^* = e^*$ on $\tilde{\mathbb{D}}$. The desired conclusion now follows by the definition of a Gaussian process on a Banach space.□

## 18.5 Exercises

**18.5.1.** Prove Proposition 18.2.

**18.5.2.** Show that Theorem 18.7 is a special case of Theorem 18.8.

**18.5.3.** Consider the sign test example discussed at the end of Section 18.3, where $T_n = \int_{\mathbb{R}} \text{sign}(x) d\mathbb{F}_n$. What value of $S_n$ should be used so that the test that rejects $H_0$ when $\sqrt{n}T_n/S_n > z_{1-\alpha}$ is asymptotically optimal at level $\alpha$?

**18.5.4.** Show that both $\phi : \mathbb{D} \mapsto \tilde{\mathbb{D}}$ and $\phi^{-1} : \tilde{\mathbb{D}} \mapsto \mathbb{D}$ as defined in the proof of Proposition 18.14 are continuous and linear.

## 18.6 Notes

Theorem 18.3 is part of Theorem 5.2.1 of BKRW, while Theorem 18.4 is their Proposition 5.2.1. Theorem 18.6 and Lemma 18.10 are Theorem 25.47 and Lemma 25.49 of van der Vaart (1998), respectively, while Theorem 18.9 is a minor modification of Theorem 25.48 of van der Vaart (1998). The proof of Lemma 18.10 given above is somewhat simpler than the corresponding proof of van der Vaart's, primarily as a result of a different utilization of linear functionals. Theorem 18.11 and Lemma 18.13 are Theorem 25.44 and Lemma 25.45 of van der Vaart (1998).

# 19

# Efficient Inference for Finite-Dimensional Parameters

In this chapter, as in Section 3.2, we focus on semiparametric models $\{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$, where $\Theta$ is an open subset of $\mathbb{R}^k$ and $H$ is an arbitrary, possibly infinite-dimensional set. The parameter of interest for this chapter is $\psi(P_{\theta,\eta}) = \theta$.

We first present the promised proof of Theorem 3.1. It may at first appear that Theorem 3.1 can only be used when the form of the efficient score equation can be written down explicitly. However, even in those settings where the efficient score cannot be written in closed form, Theorem 3.1 and a number of related approaches can be used for semiparametric efficient estimation based on the profile likelihood. This process can be facilitated through the use of approximately least-favorable submodels, which are discussed in the second section.

The main ideas of this chapter are given in Section 19.3, which presents several methods of inference for $\theta$ that go significantly beyond Theorem 3.1. The first two methods are based on a multivariate normal approximation of the profile likelihood that is valid in a neighborhood of the true parameter. The first of the two methods is based on a quadratic expansion that is valid in a shrinking neighborhood. The second method, the *profile sampler* is based on an expansion that is valid on a compact, fixed neighborhood. The third method is an extension that is valid for penalized profile likelihoods. A few other methods are also discussed, including bootstrap, jackknife, and fully Bayesian approaches.

## 19.1 Efficient Score Equations

The purpose of this section is to prove Theorem 3.1 on Page 44. Recall that this theorem was used in Section 3.2 to verify efficiency of the regression parameter estimator for the Cox model based on an estimating equation that approximated the efficient score. However, it is worth reiterating that this theorem is probably more useful for semiparametric maximum likelihood estimation, including penalized estimation, and not quite as useful as a direct method of estimation. The reason for this is that the efficient score is typically too complicated to realistically use for estimation directly, except for certain special cases like the Cox model under right censoring, and often does not have a closed form, while maximization of the semiparametric likelihood is usually much less difficult. This theorem will be utilized some in the case studies of Chapter 22.

**Proof of Theorem 3.1 (Page 44).** Define $\mathcal{F}$ to be the $P_{\theta,\eta}$-Donsker class of functions that contains both $\hat{\ell}_{\hat{\theta}_n,n}$ and $\tilde{\ell}_{\theta,\eta}$ with probability tending towards 1. By Condition (3.7), we now have

$$\mathbb{G}_n\hat{\ell}_{\hat{\theta}_n,n} = \mathbb{G}_n\tilde{\ell}_{\theta,\eta} + o_P(1) = \sqrt{n}\mathbb{P}_n\tilde{\ell}_{\theta,\eta} + o_P(1).$$

Combining this with the "no-bias" Condition (3.6), we obtain

$$\begin{aligned}\sqrt{n}(P_{\hat{\theta}_n,\eta} - P_{\theta,\eta})\hat{\ell}_{\hat{\theta}_n,n} &= o_P(1+\sqrt{n}\|\hat{\theta}_n-\theta\|) + \mathbb{G}_n\hat{\ell}_{\hat{\theta}_n,n} \\ &= \sqrt{n}\mathbb{P}_n\tilde{\ell}_{\theta,\eta} + o_P(1+\sqrt{n}\|\hat{\theta}_n-\theta\|).\end{aligned}$$

If we can show

$$\sqrt{n}(P_{\hat{\theta}_n,\eta} - P_{\theta,\eta})\hat{\ell}_{\hat{\theta}_n,n} = (\tilde{I}_{\theta,\eta} + o_P(1))\sqrt{n}(\hat{\theta}_n-\theta), \tag{19.1}$$

then the proof will be complete.

Since $\tilde{I}_{\theta,\eta} = P_{\theta,\eta}\tilde{\ell}_{\theta,\eta}\dot{\ell}'_{\theta,\eta}$,

$$\begin{aligned}&\sqrt{n}(P_{\hat{\theta}_n,\eta} - P_{\theta,\eta})\hat{\ell}_{\hat{\theta}_n,n} - \tilde{I}_{\theta,\eta}\sqrt{n}(\hat{\theta}_n-\theta) \\ &\quad= \sqrt{n}(P_{\hat{\theta}_n,\eta} - P_{\theta,\eta})\hat{\ell}_{\hat{\theta}_n,n} - P_{\theta,\eta}\left[\tilde{\ell}_{\theta,\eta}\dot{\ell}'_{\theta,\eta}\right]\sqrt{n}(\hat{\theta}_n-\theta) \\ &\quad= \sqrt{n}\int \hat{\ell}_{\hat{\theta}_n,n}(dP^{1/2}_{\hat{\theta}_n,\eta} + dP^{1/2}_{\theta,\eta}) \\ &\qquad\times\left[(dP^{1/2}_{\hat{\theta}_n,n} - dP^{1/2}_{\theta,\eta}) - \frac{1}{2}(\hat{\theta}_n-\theta)'\dot{\ell}_{\theta,\eta}dP^{1/2}_{\theta,\eta}\right] \\ &\qquad+\int\left\{\hat{\ell}_{\hat{\theta}_n,n}(dP^{1/2}_{\hat{\theta}_n,\eta} - dP^{1/2}_{\theta,\eta})\frac{1}{2}\dot{\ell}'_{\theta,\eta}dP^{1/2}_{\theta,\eta}\right\}\sqrt{n}(\hat{\theta}_n-\theta) \\ &\qquad+\int(\hat{\ell}_{\hat{\theta}_n,n} - \tilde{\ell}_{\theta,\eta})\dot{\ell}'_{\theta,\eta}dP_{\theta,\eta}\sqrt{n}(\hat{\theta}_n-\theta) \\ &\quad\equiv A_n + B_n + C_n.\end{aligned}$$

Combining the assumed differentiability in quadratic mean with Condition (3.7), we obtain $A_n = o_P(\sqrt{n}\|\hat{\theta}_n - \theta\|)$. Condition (3.7) also implies $C_n = o_P(\sqrt{n}\|\hat{\theta}_n - \theta\|)$.

Now we consider $B_n$. Note that for any sequence $m_n \to \infty$,

$$\begin{aligned}
&\left\| \int \hat{\ell}_{\hat{\theta}_n,n}(dP^{1/2}_{\hat{\theta}_n,\eta} - dP^{1/2}_{\theta,\eta})\frac{1}{2}\dot{\ell}'_{\theta,\eta}dP^{1/2}_{\theta,\eta} \right\| \\
&\qquad \leq \quad m_n \int \|\hat{\ell}_{\hat{\theta}_n,n}\| dP^{1/2}_{\theta,\eta} \left| dP^{1/2}_{\hat{\theta}_n,\eta} - dP^{1/2}_{\theta,\eta} \right| \\
&\qquad\quad + \sqrt{\int \|\hat{\ell}_{\hat{\theta}_n,n}\|^2 (dP_{\hat{\theta}_n,\eta} + dP_{\theta,\eta}) \int_{\|\dot{\ell}_{\theta,\eta}\| > m_n} \|\dot{\ell}_{\theta,\eta}\|^2 dP_{\theta,\eta}} \\
&\qquad \equiv \quad m_n D_n + E_n,
\end{aligned}$$

where $E_n = o_P(1)$ by Condition (3.7) and the square-integrability of $\dot{\ell}_{\theta,\eta}$. Now

$$\begin{aligned}
D_n^2 \quad &\leq \quad \int \|\hat{\ell}_{\hat{\theta}_n,n}\|^2 dP_{\theta,\eta} \times \int (dP^{1/2}_{\hat{\theta}_n,\eta} - dP^{1/2}_{\theta,\eta})^2 \\
&\equiv \quad F_n \times G_n,
\end{aligned}$$

where $F_n = O_P(1)$ by reapplication of Condition (3.7) and $G_n = o_P(1)$ by differentiability in quadratic mean combined with the consistency of $\hat{\theta}_n$ (see Exercise 19.5.1). Thus there exists some sequence $m_n \to \infty$ slowly enough so that $m_n^2 G_n = o_P(1)$. Hence, for this choice of $m_n$, $m_n D_n = o_P(1)$. Thus $B_n = o_P(\sqrt{n}\|\hat{\theta}_n - \theta\|)$, and we obtain (19.1). The desired result now follows.□

## 19.2 Profile Likelihood and Least-Favorable Submodels

In this section, we will focus on the maximum likelihood estimator $\hat{\theta}_n$ based on maximizing the joint empirical log-likelihood $L_n(\theta, \eta) \equiv n\mathbb{P}_n l(\theta, \eta)$, where $l(\cdot, \cdot)$ is the log-likelihood for a single observation. In this chapter, $\eta$ is regarded as a nuisance parameter, and thus we can restrict our attention to the profile log-likelihood $\theta \mapsto pL_n(\theta) \equiv \sup_\eta L_n(\theta, \eta)$. Note that $L_n$ is a sum and not an average, since we multiplied the empirical measure by $n$. In the remaining sections of the chapter, we will use a zero subscript to denote the true parameter values. While the solution of an efficient score equation need not be a maximum likelihood estimator, it is also possible that the maximum likelihood estimator in a semiparametric model may not be expressible as the zero of an efficient score equation. This possibility occurs because the efficient score is a projection, and, as such, there is no assurance that this projection is the derivative of the log-likelihood

along a submodel. To address this issue, we can utilize the approximately least-favorable submodels introduced in Section 3.3.

Recall that an approximately least-favorable submodel approximates the true least-favorable submodel (defined in the early part of Section 3.1) to a useful level of accuracy that facilitates analysis of semiparametric estimators. We will now describe this process in generality: the specifics will depend on the situation. As described in Section 3.3, we first need a general map from the neighborhood of $\theta$ into the parameter set for $\eta$, which map we will denote by $t \mapsto \eta_t(\theta, \eta)$, for $t \in \mathbb{R}^k$. We require that

$$\begin{aligned} \eta_t(\theta,\eta) &\in \hat{H}, \text{ for all } \|t-\theta\| \text{ small enough, and} \\ \eta_\theta(\theta,\eta) &= \eta \text{ for any } (\theta,\eta) \in \Theta \times \hat{H}, \end{aligned} \tag{19.2}$$

where $\hat{H}$ is a suitable enlargement of $H$ that includes all estimators that satisfy the constraints of the estimation process. Now define the map

$$\ell(t,\theta,\eta) \equiv l(t,\eta_t(\theta,\eta)).$$

We will require several things of $\ell(\cdot,\cdot,\cdot)$, at various point in our discussion, that will result in further restrictions on $\eta_t(\theta,\eta)$.

Define $\dot{\ell}(t,\theta,\eta) \equiv (\partial/(\partial t))\ell(t,\theta,\eta)$, and let $\hat{\ell}_{\theta,n} \equiv \dot{\ell}(\theta,\theta,\hat{\eta}_n)$. Clearly, $\mathbb{P}_n\hat{\ell}_{\hat{\theta}_n,n} = 0$, and thus $\hat{\theta}_n$ is efficient for $\theta_0$, provided $\hat{\ell}_{\theta,n}$ satisfies the conditions of Theorem 3.1. The reason it is necessary to check this even for maximum likelihood estimators is that often $\hat{\eta}_n$ is on the boundary (or even a little bit outside of) the parameter space. Consider for example the Cox model for right censored data (see Section 4.2.2). In this case, $\eta$ is the baseline integrated hazard function which is usually assumed to be continuous. However, $\hat{\eta}_n$ is the Breslow estimator, which is a right-continuous step function with jumps at observed failure times and is therefore not in the parameter space. Thus direct differentiation of the log-likelihood at the maximum likelihood estimator will not yield an efficient score equation.

We will also require that

$$\dot{\ell}(\theta_0,\theta_0,\eta_0) = \tilde{\ell}_{\theta_0,\eta_0}. \tag{19.3}$$

Note that we are only insisting that this identity holds at the true parameter values. We will now review four examples to illustrate this concept, the Cox model for right censored data, the proportional odds model for right censored data, the Cox model for current status data, and partly linear logistic regression. Some of these examples have already been studied extensively in previous chapters. The approximately least-favorable submodel structure we present here will be utilized later in this chapter for developing methods of inference for $\theta$.

### *19.2.1 The Cox Model for Right Censored Data*

The Cox model for right censored data has been discussed previously in this book in a number of places, including Section 4.2.2. Because of

notational tradition, we will use $\beta$ instead of $\theta$ for the regression parameter and $\Lambda$ instead $\eta$ for the baseline integrated hazard function. Recall that an observation from this model has the form $X = (W, \delta, Z)$, where $W = T \wedge C$, $\delta = 1\{W = T\}$, $Z \in \mathbb{R}^k$ is a regression covariate, $T$ is a right-censored failure time with integrated hazard given $Z$ equal to $t \mapsto e^{\beta' Z}\Lambda(t)$, and $C$ is a censoring time independent of $T$ given $Z$ and uninformative of $(\beta, \Lambda)$. We assume that there exists a $\tau < \infty$ such that $P(C \geq \tau) = P(C = \tau) > 0$. We require $H$ to consist of all monotone increasing, functions $\Lambda \in C[0, \tau]$ with $\Lambda(0) = 0$. We define $\hat{H}$ to be the set of all monotone, increasing functions $\Lambda \in D[0, \tau]$.

As shown in (3.3), the efficient score for $\beta$ is $\tilde{\ell}_{\beta,\Lambda} = \int_0^\tau (Z - h_0(s))dM(s)$, where $M(t) \equiv N(t) - \int_0^t Y(s)e^{\beta' Z}d\Lambda(s)$, $N$ and $Y$ are the usual counting and at-risk processes respectively, and

$$h_0(t) \equiv \frac{P[Z1\{W \geq t\}e^{\beta_0' Z}]}{P[1\{W \geq t\}e^{\beta_0' Z}]},$$

where $P$ is the true probability measure (at the parameter values $(\beta_0, \Lambda_0)$). Recall that the log-likelihood for a single observation is

$$l(\theta, \Lambda) = (\beta' Z + \log \Delta\Lambda(W))\delta - e^{\beta' Z}\Lambda(W),$$

where $\Delta\Lambda(w)$ is the jump size in $\Lambda$ at $w$.

If we let $t \mapsto \Lambda_t(\beta, \Lambda)) \equiv \int_0^{(\cdot)}(1 + (\beta - t)' h_0(s))d\Lambda(s)$, then $\Lambda_t(\beta, \Lambda) \in \hat{H}$ for all $t$ small enough, $\Lambda_\beta(\beta, \Lambda) = \Lambda$, and

$$\dot{\ell}(\beta_0, \beta_0, \Lambda_0) = \int_0^\tau (Z - h_0(s))dM(s) = \tilde{\ell}_{\beta_0,\Lambda_0} \tag{19.4}$$

(see Exercise 19.5.2). Thus Conditions (19.2) and (19.3) are both satisfied. We will use these results later in this chapter to develop valid methods of inference for $\beta$.

### *19.2.2 The Proportional Odds Model for Right Censored Data*

This model was studied extensively in Section 15.3. The data and parameter constraints are the same as in the Cox model, but the survival function given the covariates has the proportional odds form (15.9) instead of the proportional hazards form. Moreover, we use $A$ for the baseline integrated "hazard" instead of $\Lambda$, thus the composite parameter is $(\beta, A)$. Recall the score and information operators derived for this model from Section 15.3, and, for a fixed or random function $s \mapsto g(s)$, denote

$$W^\tau(\theta)(g) \equiv \int_0^\tau g(s)dN(s) - (1 + \delta)\frac{\int_0^{U \wedge \tau} e^{\beta' Z} g(s)dA(s)}{1 + e^{\beta' Z}A(U \wedge \tau)}.$$

Thus the usual score for $\beta$ in the direction $h_1 \in \mathbb{R}^k$, is $\dot{\ell}_{\beta,A} = V^\tau(\theta)(Z'h_1)$. We will now argue that $\tilde{\ell}_{\beta,A} = W^\tau(\theta)\left(Z'h_1 - [\sigma_\theta^{22}]^{-1}\sigma_\theta^{21}(h_1)(\cdot)\right)$, where $\sigma_\theta^{22}$ is continuously invertible, and thus the least favorable direction for this model is

$$s \mapsto h_0(s) \equiv [\sigma_\theta^{22}]^{-1}\sigma_\theta^{21}(h_1)(s).$$

Recall that the operator $\sigma_\theta^{21}$ is compact, linear operator from $\mathbb{R}^k$ to $\mathcal{H}_\infty^2$ $\equiv$ "the space of functions on $D[0,\tau]$ with bounded total variation", and the operator $\sigma_\theta^{22}$ is a linear operator from $\mathcal{H}_\infty^2$ to $\mathcal{H}_\infty^2$. We first need to show that $\sigma_\theta^{22}$ is continuously invertible, and then we need to show that $h_1 \mapsto W^\tau(h_0'(\cdot)h_1)$ is the projection of $h_1 \mapsto W^\tau(Z'h_1)$ onto the closed linear span of $h_2 \mapsto V^\tau(h_2(\cdot))$, where $h_2$ ranges over $\mathcal{H}_\infty^2$. The first result follows from the same arguments used in the proof of Theorem 15.9, but under a simplified model where $\beta$ is known to have the value $\beta_0$ (see Exercise 19.5.3).

For the second result, it is clear based on the first result that the function $s \mapsto [\sigma_\theta^{22}]^{-1}\sigma_\theta^{21}(h_1)(s)$ is an element of $\mathcal{H}_\infty^2$ for every $h_1 \in \mathbb{R}^k$. Thus $W^\tau(\theta)(h_0'(\cdot)h_1)$ is in the closed linear span of $W^\tau(\theta)(h_2)$, where $h_2$ ranges over $\mathcal{H}_\infty^2$, and thus all that remains to verify is that

$$P\left[\left\{W^\tau(\theta)(Z'h_1) - W^\tau(\theta)(h_0'(\cdot)h_1)\right\} W^\tau(\theta)(h_2)\right] = 0, \text{ for all } h_2 \in \mathcal{H}_\infty^2. \tag{19.5}$$

Considering the relationship between $W^\tau$ and the $V^\tau$ defined in Section 15.3.4, we have that the left side of (19.5) equals

$$\begin{aligned} & P\left[V^\tau(\theta)\begin{pmatrix} Z'h_1 \\ -[\sigma_\theta^{22}]^{-1}\sigma_\theta^{21}(h_1) \end{pmatrix} V^\tau(\theta)\begin{pmatrix} 0 \\ h_2 \end{pmatrix}\right] \\ &= h_1'\sigma_\theta^{11}(0) + h_1'\sigma_\theta^{12}(h_2) \\ &\quad -(0)\sigma_\theta^{12}[\sigma_\theta^{22}]^{-1}\sigma_\theta^{21}(h_1) - \int_0^\tau h_2(s)\sigma_{21}(h_1)(s)dA(s) \\ &= 0, \end{aligned}$$

since

$$h_1'\sigma_\theta^{12}(h_2) = \int_0^\tau h_2(s)\sigma_{21}(h_1)(s)dA(s),$$

as can be deduced from the definitions following (15.20). Since $h_2$ was arbitrary, we have proven (19.5). Thus $h_0 \in \mathcal{H}_\infty^2$ as defined previously in this section is indeed the least favorable direction. Hence $t \mapsto A_t(\beta, A) \equiv \int_0^{(\cdot)}(1 + (\beta - t)'h_0(s))dA(s)$ satisfies both (19.2) and (19.3).

### *19.2.3 The Cox Model for Current Status Data*

Current status data arises when each subject is observed at a single examination time, $Y$, to determine whether an event has occurred. The event time, $T$, cannot be observed exactly. Including the covariate $Z$, the observed data in this setting consists of $n$ independent and identically distributed realizations of $X = (Y, \delta, Z)$, where $\delta = 1\{T \leq Y\}$. We assume that the integrated hazard function of $T$ given $Z$ has the proportional hazards form and parameters $(\beta, \Lambda)$ as given in Section 19.2.1 above.

We also make the following additional assumptions. $T$ and $Y$ are independent given $Z$. $Z$ lies in a compact set almost surely and the covariance of $Z - E(Z|Y)$ is positive definite, which, as we will show later, guarantees that the efficient information $\tilde{I}_0$ is strictly positive. $Y$ possesses a Lebesgue density which is continuous and positive on its support $[\sigma, \tau]$, where $0 < \sigma < \tau < \infty$, for which the true parameter $\Lambda_0$ satisfies $\Lambda_0(\sigma-) > 0$ and $\Lambda_0(\tau) < M < \infty$, for some known $M$, and is continuously differentiable on this interval with derivative bounded above zero. We let $H$ denote all such possible choices of $\Lambda_0$ satisfying these constraints for the given value of $M$, and we let $\hat{H}$ consist of all nonnegative, nondecreasing right-continuous functions $\Lambda$ on $[\sigma, \tau]$ with $\Lambda(\tau) \leq M$.

We can deduce that the log-likelihood for a single observation, $l(\beta, \Lambda)$, has the form

$$l(\beta, \Lambda) = \delta \log[1 - \exp(-\Lambda(Y)\exp(\beta' Z))] - (1-\delta)\exp(\beta' Z)\Lambda(Y). \tag{19.6}$$

From this, we can deduce that the score function for $\beta$ takes the form $\dot{\ell}_{\beta,\Lambda}(x) = z\Lambda(y)Q(x;\beta,\Lambda)$, where

$$Q(x;\beta,\Lambda) = e^{\beta' z}\left[\delta \frac{\exp(-e^{\beta' z}\Lambda(y))}{1 - \exp(-e^{\beta' z}\Lambda(y))} - (1-\delta)\right].$$

Inserting a submodel $t \mapsto \Lambda_t$ such that $h(y) = -\partial/\partial t|_{t=0}\Lambda_t(y)$ exists for every $y$ into the log likelihood and differentiating at $t = 0$, we obtain a score function for $\Lambda$ of the form $A_{\beta,\Lambda}h(x) = h(y)Q(x;\beta,\Lambda)$. The linear span of these functions contains $A_{\beta,\Lambda}h$ for all bounded functions $h$ of bounded variation. Thus the efficient score function for $\theta$ is $\tilde{\ell}_{\beta,\Lambda} = \dot{\ell}_{\beta,\Lambda} - A_{\beta,\Lambda}h_{\beta,\Lambda}$ for the least-favorable direction vector of functions $h_{\beta,\Lambda}$ minimizing the distance $P_{\beta,\Lambda}\|\dot{\ell}_{\beta,\Lambda} - A_{\beta,\Lambda}h\|^2$. The solution at the true parameter is

$$\begin{aligned} h_0(Y) &\equiv \Lambda_0(Y)h_{00}(Y) \\ &\equiv \Lambda_0(Y)\frac{E_{\beta_0,\Lambda_0}(ZQ^2(X;\beta_0,\Lambda_0)|Y)}{E_{\beta_0,\Lambda_0}(Q^2(X;\beta_0,\Lambda_0)|Y)} \end{aligned} \tag{19.7}$$

(see Exercise 19.5.4). As the formula shows, the vector of functions $h_0(y)$ is unique a.s., and $h_0(y)$ is a bounded function since $Q(x;\theta_0,\Lambda_0)$ is bounded

away from zero and infinity. We assume that the function $h_0$ given by (19.7) has a version which is differentiable with a bounded derivative on $[\sigma, \tau]$.

An approximately least-favorable submodel can thus be of the form $\Lambda_t(\beta, \Lambda)(\cdot) = \Lambda(\cdot) + \phi(\Lambda(\cdot))(\beta - t)' h_{00} \circ \Lambda_0^{-1} \circ \Lambda(\cdot)$, where $\phi(\cdot)$ is a fixed function we will define shortly that approximates the identity. Note that we need to extend $\Lambda_0^{-1}$ so that it is defined on all of $[0, M]$. This is done by letting $\Lambda_0^{-1}(t) = \sigma$ for all $t \in [0, \Lambda_0(\sigma)]$ and $\Lambda_0^{-1}(t) = \tau$ for all $t \in [\Lambda_0(\tau), M]$. We take $\phi : [0, M] \mapsto [0, M]$ to be any fixed function with $\phi(u) = u$ on $[\Lambda_0(\sigma), \Lambda_0(\tau)]$ such that $u \mapsto \phi(u)/u$ is Lipschitz and $\phi(u) \leq c(u \wedge (M - u))$ for all $u \in [0, M]$ and some $c < \infty$ depending only on $(\beta_0, \Lambda_0)$. Our conditions on the model ensure that such a function exists.

The function $\Lambda_t(\beta, \Lambda)$ is essentially $\Lambda$ plus a perturbation in the least favorable direction, $h_0$, but its definition is somewhat complicated in order to ensure that $\Lambda_t(\beta, \Lambda)$ really defines a cumulative hazard function within our parameter space, at least for all $t$ that is sufficiently close to $\beta$. To see this, first note that by using $h_{00} \circ \Lambda_0^{-1} \circ \Lambda$, rather than $h_{00}$, we ensure that the perturbation that is added to $\Lambda$ is Lipschitz-continuous with respect to $\Lambda$. Combining this with the Lipschitz-continuity of $\phi(u)/u$, we obtain that for any $0 \leq v < u \leq M$,

$$\Lambda_t(\beta, \Lambda)(u) - \Lambda_t(\beta, \Lambda)(v) \geq \Lambda(u) - \Lambda(v) - \Lambda(v) t k_0 (\Lambda(u) - \Lambda(v)),$$

for some universal constant $0 < k_0 < \infty$. Since $\Lambda(v) \leq M$, we obtain that for all $\|t - \beta\|$ small enough, $\Lambda_t(\beta, \Lambda)$ is non-decreasing. The additional constraints on $\phi$ ensures that $0 \leq \Lambda_t(\beta, \Lambda) < M$ for all $\|t - \beta\|$ small enough.

Hence $t \mapsto \Lambda_t(\beta, \Lambda) \equiv \int_0^{(\cdot)} (1 + (\beta - t)' h_0(s)) d\Lambda(s)$ satisfies both (19.2) and (19.3). Additional details about the construction of the approximately least-favorable submodel for this example can be found in Section 4.1 of Murphy and van der Vaart (2000).

### *19.2.4 Partly Linear Logistic Regression*

This a continuation of the example discussed in Chapter 1 and Sections 4.5 and 15.1. Recall that an observation from this model has the form $X = (Y, Z, U)$, where $Y$ is a dichotomous outcome, $Z \in \mathbb{R}^d$ is a covariate, and $U \in [0, 1]$ is an additional, continuous covariate. Moreover, the probability that $Y = 1$ given that $(Z, U) = (z, u)$ is $\nu[\beta' z + \eta(u)]$, where $u \mapsto \nu(u) \equiv e^u/(1 + e^u)$, $\eta \in \mathcal{H}$, and $\mathcal{H}$ consists of those functions $h \in C[0, 1]$ with $J(h) \equiv \int_0^1 [h^{(d)}(s)]^2 ds < \infty$, where $h^{(j)}$ is the $j$th derivative of $h$, and $d < \infty$ is known, positive integer. Additional assumptions are given in Sections 4.5 and 15.1. These assumptions include $Z$ being restricted to a known, compact set, and the existence of a known $c_0 < \infty$ for which $\|\beta\| < c_0$ and $\|h\|_\infty < c_0$. Thus $H = \{h \in \mathcal{H} : \|h\|_\infty < c_0\}$ and $\hat{H} = \mathcal{H}$.

As shown in Section 4.5, the efficient score for $\beta$ is $\tilde{\ell}_{\beta,\eta}(Y,Z,U) \equiv (Z - h_1(Y))(Y - \mu_{\beta,\eta}(Z,U))$, where $\mu_{\beta,\eta}(Z,U) \equiv \nu[\beta' Z + \eta(U)]$,

$$h_1(u) \equiv \frac{\mathrm{E}\{ZV_{\beta,\eta}(Z,U)|U=u\}}{\mathrm{E}\{V_{\beta,\eta}(Z,U)|U=u\}},$$

and $V_{\beta,\eta}(Z,U) \equiv \mu_{\beta,\eta}(Z,U)(1-\mu_{\beta,\eta}(Z,U))$. Thus $h_1$ is the least-favorable direction, and, provided there exists a version of $h_1$ such that $h_1 \in \hat{H}$, then $\eta_t(\beta,\eta) = \eta + (\beta - t)' h_1$ satisfies both (19.2) and (19.3). This example differs from the previous examples since the likelihood used for estimation is penalized.

## 19.3 Inference

We now discuss a few methods of conducting efficient estimation and inference for $\theta$ using the profile likelihood. The first method is based on the very important result that, under reasonable regularity conditions, a profile likelihood for $\theta$ behaves asymptotically like a parametric likelihood of a normal random variable with variance equal to the inverse of the efficient Fisher information $\tilde{I}_{\theta,\eta}$ in a shrinking neighborhood of the maximum likelihood estimator $\theta_0$. This can yield valid likelihood ratio based inference for $\theta$. The second method, the *profile sampler*, extends the first method to show that this behavior occurs in a compact neighborhood of $\theta_0$, and thus an automatic method for estimation and inference can be obtain by applying a simple random walk on the profile likelihood. The third method extends the idea of the second method to penalized maximum likelihoods. Several other methods, including the bootstrap, jackknife and purely Bayesian approaches, are also discussed.

### *19.3.1 Quadratic Expansion of the Profile Likelihood*

The main ideas of this section come from a very elegant paper by Murphy and van der Vaart (2000) on profile likelihood. The context is the same as in the beginning of Section 19.2, wherein we have maximum likelihood estimators $(\hat{\theta}_n, \hat{\eta}_n)$ based on a i.i.d. sample $X_1, \ldots, X_n$, where one is the finite-dimensional parameter of primary interest $(\hat{\theta}_n)$ and the other is an infinite-dimensional nuisance parameter $(\hat{\eta}_n)$. The main result, given formally below as Theorem 19.5 (on Page 359), is that under certain regularity conditions, we have for any estimator $\tilde{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$, that

$$
\begin{aligned}
(19.8) \qquad pL_n(\tilde{\theta}_n) &= pL_n(\theta_0) + (\tilde{\theta}_n - \theta_0)' \sum_{i=1}^{n} \tilde{\ell}_{\theta_0,\eta_0}(X_i) \\
&\quad -\frac{1}{2} n (\tilde{\theta}_n - \theta_0)' \tilde{I}_{\theta_0,\eta_0} (\tilde{\theta}_n - \theta_0) \\
&\quad +o_{P_0}(1 + \sqrt{n}\|\tilde{\theta}_n - \theta_0\|)^2,
\end{aligned}
$$

where $\tilde{I}_{\theta_0,\eta_0}$ is the efficient Fisher information and $P_0$ the probability measure of $X$ at the true parameter values.

Suppose we can know that the maximum profile likelihood estimator is consistent, i.e., that $\hat{\theta}_n = \theta_0 + o_{P_0}(1)$, and that $\tilde{I}_{\theta_0,\eta_0}$ is positive definite. Then if (19.8) also holds, we have that

$$
\begin{aligned}
\|\sqrt{n}(\hat{\theta}_n - \theta_0)\|^2 &\leq \sqrt{n}(\hat{\theta}_n - \theta_0)' \left[ n^{-1/2} \sum_{i=1}^{n} \tilde{\ell}_{\theta_0,\eta_0}(X_i) \right] \\
&\quad +o_{P_0}(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|)^2 \\
&= O_{P_0}(\sqrt{n}\|\hat{\theta}_n - \theta_0\|) + o_{P_0}(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|)^2,
\end{aligned}
$$

since $pL_n(\hat{\theta}_n) - pL_n(\theta_0) \geq 0$. This now implies that

$$
(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|)^2 = O_{P_0}(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|) + o_{P_0}(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|)^2,
$$

which yields that $\sqrt{n}\|\hat{\theta}_n - \theta_0\| = O_{P_0}(1)$.

Let $K \subset \mathbb{R}^k$ be a compact neighborhood of 0, and note that for any sequence of possibly random points $\theta_n \in \theta_0 + n^{-1/2}K$, we have $\theta_n = \theta_0 + o_{P_0}(1)$ and $\sqrt{n}(\theta_n - \theta_0) = O_{P_0}(1)$. Thus $\sup_{u \in K} |pL_n(\theta_0 + u/\sqrt{n}) - pL_n(\theta_0) - M_n(u)| = o_{P_0}(1)$, where $u \mapsto M_n(u) \equiv u'Z_n - (1/2)u'\tilde{I}_{\theta_0,\eta_0}u$ and $Z_n \equiv \sqrt{n}\mathbb{P}_n\tilde{\ell}_{\theta_0,\eta_0}(X)$. Since $Z_n \rightsquigarrow Z$, where $Z \sim N_k(0, \tilde{I}_{\theta_0,\eta_0})$, and since $K$ was arbitrary, we have by the argmax theorem (Theorem 14.1) that $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \tilde{I}^{-1}_{\theta_0,\eta_0} Z$, which implies that $\hat{\theta}_n$ is efficient, and thus, by Theorem 18.7, we also have

$$
(19.9) \qquad \sqrt{n}(\hat{\theta}_n - \theta_0) = \sqrt{n}\mathbb{P}_n \tilde{I}^{-1}_{\theta_0,\eta_0} \tilde{\ell}_{\theta_0,\eta_0}(X) + o_{P_0}(1).
$$

These arguments can easily be strengthened to imply the following simple corollary:

COROLLARY 19.1 *Let the estimator $\check{\theta}_n$ be consistent for $\theta_0$ and satisfy $pL_n(\check{\theta}_n) \geq pL_n(\hat{\theta}_n) - o_{P_0}(1)$. Then, provided $\tilde{I}_{\theta_0,\eta_0}$ is positive definite and (19.8) holds, $\check{\theta}_n$ is efficient.*

The proof is not difficult and is saved as an exercise (see Exercise 19.5.5).

Combining (19.8) and Corollary 19.1, we obtain after some algebra the following profile likelihood expansion centered around any consistent, approximate maximizer of the profile likelihood:

COROLLARY 19.2 *Let $\check{\theta}_n = \theta_0 + o_{P_0}(1)$ and satisfy $pL_n(\check{\theta}_n) \geq pL_n(\hat{\theta}_n) - o_{P_0}(1)$. Then, provided $\tilde{I}_{\theta_0,\eta_0}$ is positive definite and (19.8) holds, we have for any random sequence $\tilde{\theta}_n = \theta_0 + o_{P_0}(1)$,*

$$pL_n(\tilde{\theta}_n) = pL_n(\check{\theta}_n) - \frac{1}{2}n(\tilde{\theta}_n - \check{\theta}_n)'\tilde{I}_{\theta_0,\eta_0}(\tilde{\theta}_n - \check{\theta}_n) + o_{P_0}(1 + \sqrt{n}\|\tilde{\theta}_n - \theta_0\|)^2. \tag{19.10}$$

The proof is again saved as an exercise (see Exercise 19.5.6).

The following two additional corollaries provide methods of using this quadratic expansion to conduct inference for $\theta_0$:

COROLLARY 19.3 *Assume the conditions of Corollary 19.1 hold for $\check{\theta}_n$. Then, under the null hypothesis $H_0 : \theta = \theta_0$, $2(pL_n(\check{\theta}_n) - pL_n(\theta_0)) \rightsquigarrow \chi^2(k)$, where $\chi^2(k)$ is a chi-squared random variable with $k$ degrees of freedom.*

COROLLARY 19.4 *Assume the conditions of Corollary 19.1 hold for $\check{\theta}_n$. Then for any vector sequence $v_n \xrightarrow{\mathrm{P}} v \in \mathbb{R}^k$ and any scalar sequence $h_n \xrightarrow{\mathrm{P}} 0$ such that $(\sqrt{n}h_n)^{-1} = O_P(1)$, where the convergence is under $P = P_0$, we have*

$$-2\frac{pL_n(\check{\theta}_n + h_n v_n) - pL_n(\check{\theta}_n)}{nh_n^2} \xrightarrow{\mathrm{P}} v'\tilde{I}_{\theta_0,\eta_0}v.$$

**Proofs.** Corollary 19.3 follows immediately from Corollary 19.2 by setting $\tilde{\theta}_n = \theta_0$, while Corollary 19.4 also follows from Corollary 19.2 but after setting $\tilde{\theta}_n = \check{\theta}_n + h_n v_n$.□

Corollary 19.3 can be used for hypothesis testing and confidence region construction for $\theta_0$, while Corollary 19.4 can be used to obtain consistent, numerical estimates of $\tilde{I}_{\theta_0,\eta_0}$. The purpose of the remainder of this section is to present and verify reasonable regularity conditions for (19.8) to hold. After the presentation and proof of Theorem 19.5, we will verify the conditions of the theorem for two of the examples mentioned above, the Cox model for right-censored data and the Cox model for current status data. The next section will show that slightly stronger assumptions can yield even more powerful methods of inference.

To begin with, we will need an approximately least-favorable submodel $t \mapsto \eta_t(\theta, \eta)$ that satisfies Conditions (19.2) and (19.3). Define $\ddot{\ell}(t, \theta, \eta) \equiv (\partial/(\partial t))\dot{\ell}(t, \theta, \eta)$ and $\hat{\eta}_\theta \equiv \mathrm{argmax}_\eta L_n(\theta, \eta)$, and assume that for any possibly random sequence $\tilde{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$, we have

$$\hat{\eta}_{\tilde{\theta}_n} \xrightarrow{\mathrm{P}} \eta \text{ and} \tag{19.11}$$

$$P_0\dot{\ell}(\theta_0, \tilde{\theta}_n, \hat{\eta}_{\tilde{\theta}_n}) = o_{P_0}(\|\tilde{\theta}_n - \theta_0\| + n^{-1/2}). \tag{19.12}$$

We are now ready to present the theorem:

THEOREM 19.5 *Assume Conditions (19.2), (19.3), (19.11) and (19.12) are satisfied, and assume that the functions $(t, \theta, \eta) \mapsto \dot{\ell}(t, \theta, \eta)(X)$ and*

*$(t,\theta,\eta) \mapsto \ddot{\ell}(t,\theta,\eta)(X)$ are continuous at $(\theta_0,\theta_0,\eta_0)$ for $P_0$-almost every $X$. Also assume that for some neighborhood $V$ of $(\theta_0,\theta_0,\eta_0)$, the class of functions $\mathcal{F}_1 \equiv \{\dot{\ell}(t,\theta,\eta) : (t,\theta,\eta) \in V\}$ is $P_0$-Donsker with square-integrable envelope function and the class of functions $\mathcal{F}_2 \equiv \{\ddot{\ell}(t,\theta,\eta) : (t,\theta,\eta) \in V\}$ is $P_0$-Glivenko-Cantelli and bounded in $L_1(P_0)$. Then (19.8) holds.*

**Proof.** The proof follows closely the proof of Theorem 1 given in Murphy and van der Vaart (2000). Since $\dot{\ell}(t,\theta,\eta) \to \tilde{\ell}_{\theta_0,\eta_0} \equiv \tilde{\ell}_0$ as $(t,\theta,\eta) \to (\theta_0,\theta_0,\eta_0)$, and since the envelope of $\mathcal{F}_1$ is square-integrable, we have by the dominated convergence theorem that $P_0\left(\dot{\ell}(\tilde{t},\tilde{\theta},\tilde{\eta}) - \tilde{\ell}_0\right)^2 \xrightarrow{\mathrm{P}} 0$ for every $(\tilde{t},\tilde{\theta},\tilde{\eta}) \xrightarrow{\mathrm{P}} (\theta_0,\theta_0,\eta_0)$. We likewise have $P_0\ddot{\ell}(\tilde{t},\tilde{\theta},\tilde{\eta}) \xrightarrow{\mathrm{P}} P_0\ddot{\ell}(\theta_0,\theta_0,\eta_0)$. Because $t \mapsto \exp[\ell(t,\theta_0,\eta_0)]$ corresponds to a smooth parametric likelihood up to a multiplicative constant for all $\|t-\theta_0\|$ small enough, we know that its derivatives satisfy the usual identity:

$$\begin{aligned} P_0\ddot{\ell}(\theta_0,\theta_0,\eta_0) &= -P_0\left[\dot{\ell}(\theta_0,\theta_0,\eta_0)\dot{\ell}'(\theta_0,\theta_0,\eta_0)\right] \\ &= -\tilde{I}_{\theta_0,\eta_0} \equiv -\tilde{I}_0. \end{aligned} \tag{19.13}$$

Combining these facts with the empirical process conditions, we obtain that for every $(\tilde{t},\tilde{\theta},\tilde{\eta}) \xrightarrow{\mathrm{P}} (\theta_0,\theta_0,\eta_0)$, both

$$\mathbb{G}_n\dot{\ell}(\tilde{t},\tilde{\theta},\tilde{\eta}) = \mathbb{G}_n\tilde{\ell}_0 + o_{P_0}(1) \text{ and} \tag{19.14}$$

$$\mathbb{P}_n\ddot{\ell}(\tilde{t},\tilde{\theta},\tilde{\eta}) \xrightarrow{\mathrm{P}} -\tilde{I}_0. \tag{19.15}$$

By (19.2) and the definition of $\hat{\eta}_\theta$, we have

$$n^{-1}\left[pL_n(\tilde{\theta}) - pL_n(\theta_0)\right] = \mathbb{P}_n l(\tilde{\theta},\hat{\eta}_{\tilde{\theta}}) - \mathbb{P}_n l(\theta_0,\hat{\eta}_{\theta_0}),$$

which is bounded below by $\mathbb{P}_n l(\tilde{\theta},\eta_{\tilde{\theta}}(\theta_0,\hat{\eta}_{\theta_0})) - \mathbb{P}_n l(\theta_0,\eta_{\theta_0}(\theta_0,\hat{\eta}_{\theta_0}))$ and bounded above by $\mathbb{P}_n l(\tilde{\theta},\eta_{\tilde{\theta}}(\tilde{\theta},\hat{\eta}_{\tilde{\theta}})) - \mathbb{P}_n l(\theta_0,\eta_{\theta_0}(\tilde{\theta},\hat{\eta}_{\tilde{\theta}}))$. The lower bound follows from the facts that $\hat{\eta}_{\tilde{\theta}}$ maximizes $\eta \mapsto L_n(\tilde{\theta},\eta)$ and $\eta_{\theta_0}(\theta_0,\hat{\eta}_{\theta_0}) = \hat{\eta}_{\theta_0}$, while the upper bound follows from the facts that $\eta_{\tilde{\theta}}(\tilde{\theta},\hat{\eta}_{\tilde{\theta}}) = \hat{\eta}_{\tilde{\theta}}$ and $\hat{\eta}_{\theta_0}$ maximizes $\eta \mapsto L_n(\theta_0,\eta)$. Note that both the upper and lower bounds have the form $\mathbb{P}_n\ell(\tilde{\theta},\tilde{\psi}) - \mathbb{P}_n\ell(\theta_0,\tilde{\psi})$, where $\tilde{\psi} = (\theta_0,\hat{\eta}_{\theta_0})$ for the lower bound and $\tilde{\psi} = (\tilde{\theta},\hat{\eta}_{\tilde{\theta}})$ for the upper bound. The basic idea of the remainder of the proof is to apply a two-term Taylor expansion to both the upper and lower bounds and show that they agree up to a term of order $o_{P_0}(\|\tilde{\theta}-\theta_0\| + n^{-1/2})^2$.

For some $\tilde{t}$ on the line segment between $\tilde{\theta}$ and $\theta_0$, we have that

$$\begin{aligned} \mathbb{P}_n\ell(\tilde{\theta},\tilde{\psi}) - \mathbb{P}_n\ell(\theta_0,\tilde{\psi}) &= (\tilde{\theta}-\theta_0)'\mathbb{P}_n\dot{\ell}(\theta_0,\tilde{\psi}) \\ &\quad + \frac{1}{2}(\tilde{\theta}-\theta_0)'\mathbb{P}_n\ddot{\ell}(\tilde{t},\tilde{\psi})(\tilde{\theta}-\theta_0). \end{aligned}$$

For the second term on the right, we can, as a consequence of (19.15), replace $\mathbb{P}_n\ddot{\ell}(\tilde{t},\tilde{\psi})$ with $-\tilde{I}_0$ after adding an $o_{P_0}(\|\tilde{\theta}-\theta_0\|^2)$ term. For the first term, we can utilize (19.14) to obtain that

$$\begin{aligned}\sqrt{n}\mathbb{P}_n\dot{\ell}(\theta_0,\tilde{\psi}) &= \mathbb{G}_n\tilde{\ell}_0+\mathbb{G}_n\left[\dot{\ell}(\theta_0,\tilde{\psi})-\tilde{\ell}_0\right]+\sqrt{n}P_0\dot{\ell}(\theta_0,\tilde{\psi})\\ &= \mathbb{G}_n\tilde{\ell}_0+o_{P_0}(1)+o_{P_0}(1+\sqrt{n}\|\tilde{\theta}-\theta_0\|),\end{aligned}$$

where the error terms follow from the empirical process assumptions and (19.12), respectively. Thus the first term becomes

$$(\tilde{\theta}-\theta_0)'\mathbb{P}_n\tilde{\ell}_0+o_{P_0}(\|\tilde{\theta}-\theta_0\|^2+n^{-1/2}\|\tilde{\theta}-\theta_0\|).$$

The desired conclusion now follows.□

We now verify the conditions of this theorem for two models, the Cox model for right censored data and the Cox model for current status data. The verification of the conditions for the proportional odds model under right-censoring will be considered in the case studies of Chapter 22. Inference for the partly-linear logistic regression example will be considered later in Section 19.3.3. Several other very interesting semiparametric examples for which Theorem 19.5 is applicable are presented in Murphy and van der Vaart (2000), including a case-control set-up with a missing covariate, a shared gamma-frailty model for right-censored data, and an interesting semiparametric mixture model.

*The Cox model for right censored data.*

This is a continuation of Section 19.2.1. Recall that Conditions (19.2) and (19.3) have already been established for this model. Putting the previous results for this example together, we first obtain that

$$\ell(t,\beta,\Lambda)=(\beta'Z+\log\left[\Delta\Lambda_t(\beta,\Lambda)(W)\right]\delta-e^{\beta'Z}\Lambda_t(\beta,\Lambda)(W).$$

Differentiating twice with respect to $t$, we next obtain

$$\begin{aligned}\dot{\ell}(t,\beta,\Lambda) &= Z\delta-Ze^{t'Z}\Lambda_t(\beta,\Lambda)(W)\\ &\quad -\frac{h_0(W)\delta}{1+(\beta-t)'h_0(W)}+e^{t'Z}\int_0^W h_0(s)d\Lambda(s)\end{aligned}$$

and

$$\begin{aligned}\ddot{\ell}(t,\beta,\Lambda) &= -ZZ'e^{t'Z}\Lambda_t(\beta,\Lambda)(W)+e^{t'Z}\int_0^W(Zh_0'(s)+h_0(s)Z')d\Lambda(s)\\ &\quad -\frac{h_0(W)h_0'(W)\delta}{\left[1+(\beta-t)'h_0(W)\right]^2}.\end{aligned}$$

In this case, the maximizer of $\Lambda\mapsto L_n(\beta,\Lambda)$ has the closed form of the Breslow estimator:

$$\hat{\Lambda}_\beta(s) = \int_0^s \frac{\mathbb{P}_n dN(u)}{\mathbb{P}_n Y(u) e^{\beta' Z}},$$

and it is easy to verify that $\hat{\Lambda}_{\tilde{\beta}_n}$ is uniformly consistent for $\Lambda_0$ for any possibly random sequence $\tilde{\beta}_n \xrightarrow{\mathrm{P}} \beta_0$. Thus Condition (19.11) holds. Recall the counting process notation $Y(s) \equiv 1\{W \geq s\}$ and $N(s) \equiv 1\{W \leq s\}\delta$, and note that $P_0 \dot{\ell}(\beta_0, \tilde{\beta}_n, \hat{\Lambda}_{\tilde{\beta}_n})$

$$\begin{aligned}
&= P_0 \left[ Z\delta - \int_0^\tau (1 + (\tilde{\beta}_n - \beta_0)' h_0(s)) Z Y(s) e^{\beta_0' Z} d\hat{\Lambda}_{\tilde{\beta}_n}(s) \right. \\
&\quad \left. - \frac{h_0(W)\delta}{1 + (\tilde{\beta}_n - \beta_0)' h_0(W)} + \int_0^\tau h_0(s) Y(s) e^{\beta_0' Z} d\hat{\Lambda}_{\tilde{\beta}_n}(s) \right] \\
&= P_0 \left[ - \int_0^\tau (\tilde{\beta}_n - \beta_0)' h_0(s) Z Y(s) e^{\beta_0' Z} d\hat{\Lambda}_{\tilde{\beta}_n}(s) \right. \\
&\quad \left. + \frac{h_0(W)(\tilde{\beta}_n - \beta_0)' h_0(W))\delta}{1 + (\tilde{\beta}_n - \beta_0)' h_0(W)} \right] \\
&= P_0 \left[ - \int_0^\tau Z (\tilde{\beta}_n - \beta_0)' h_0(s) Y(s) e^{\beta_0' Z} \left( d\hat{\Lambda}_{\tilde{\beta}_n}(s) - d\Lambda_0(s) \right) \right. \\
&\quad \left. + \frac{h_0(W) \left( (\tilde{\beta}_n - \beta_0)' h_0(W) \right)^2 \delta}{1 + (\tilde{\beta}_n - \beta_0)' h_0(W)} \right] \\
&= o_{P_0}(\|\tilde{\beta}_n - \beta_0\|).
\end{aligned}$$

The second equality follows from the identity

$$P_0 \left[ (Z - h_0(s)) Y(s) e^{\beta_0' Z} \right] = 0, \quad \text{for all } s \in [0, \tau], \tag{19.16}$$

while the third equality follows from the fact that

$$M(s) \equiv N(s) - \int_0^s Y(u) e^{\beta_0' Z} dN(u)$$

is a mean-zero martingale combined with a reapplication of (19.16). Thus Condition (19.12) holds.

The smoothness of the functions involved, combined with the fact that $u \mapsto h_0(u)$ is bounded in total variation, along with other standard arguments, yields the required continuity, Donsker and Glivenko-Cantelli conditions of Theorem 19.5 (see Exercise 19.5.7). Thus all of the conditions of Theorem 19.5 are satisfied for the Cox model for right censored data.

*The Cox model for current status data.*

This is a continuation of Section 19.2.3 in which we formulated an approximately least-favorable submodel $\Lambda_t(\beta, \Lambda)$ that satisfies Conditions (19.2) and (19.3). We also have from Section 19.2.3 that

$$\begin{aligned}\ell(t,\beta,\Lambda) &= l(t,\Lambda_t(\beta,\Lambda)) \\ &= \delta \log\left[1-\exp\left(-\Lambda_t(\beta,\Lambda)(Y)e^{t'Z}\right)\right] \\ &\quad -(1-\delta)e^{t'Z}\Lambda_t(\beta,\Lambda)(Y).\end{aligned}$$

This yields

$$\begin{aligned}\dot{\ell}(t,\beta,\Lambda) &= (Z\Lambda_t(\beta,\Lambda)(Y)+\dot{\Lambda}(Y))Q(X;t,\Lambda_t(\beta,\Lambda)), \text{ and} \\ \ddot{\ell}(t,\beta,\Lambda) &= \left(\Lambda_t(\beta,\Lambda)(Y)ZZ'+\dot{\Lambda}(Y)Z+Z\dot{\Lambda}'(Y)\right)Q(X;t,\Lambda_t(\beta,\Lambda)) \\ &\quad -e^{2t'Z}\left[Z\Lambda_t(\beta,\Lambda)(Y)+\dot{\Lambda}(Y)\right]^{\otimes 2} \\ &\quad \times\frac{\exp\left(-e^{t'Z}\Lambda_t(\beta,\Lambda)(Y)\right)}{\left[1-\exp\left(-e^{t'Z}\Lambda_t(\beta,\Lambda)(Y)\right)\right]^2},\end{aligned}$$

where

$$\dot{\Lambda}(u) \equiv -\phi(\Lambda(u))h_{00}\circ\Lambda_0^{-1}\circ\Lambda(u).$$

Note that these functions are smooth enough to satisfy the continuity conditions of Theorem 19.5.

Using entropy calculations, Murphy and van der Vaart (1999), extend earlier results of Huang (1996) to obtain

$$P_0\left[\hat{\Lambda}_{\tilde{\beta}_n}(Y)-\Lambda_0(Y)\right]^2 = O_{P_0}(\|\tilde{\beta}_n-\beta_0\|^2+n^{-2/3}),$$

for any possibly random sequence $\tilde{\beta}_n \xrightarrow{\mathrm{P}} \beta_0$. This yields that Condition (19.11) is satisfied. The rate of convergence also helps in verifying Condition (19.12). We omit the details in verifying (19.12), but they can be found on Page 460 of Murphy and van der Vaart (2000).

It is not difficult to verify that the carefully formulated structure of $\Lambda_t(\beta,\Lambda)$ ensures the existence of a neighborhood $V$ of $(\beta_0,\beta_0,\Lambda_0)$ for which the class $\mathcal{G} \equiv \{\Lambda_t(\beta,\Lambda) : (t,\beta,\Lambda) \in V\}$, viewed as a collection of functions of $Y$, is a subset of the class of all monotone functions $f : [0,\tau] \mapsto [0,M]$, which class is known to be Donsker for any probability measure. Since $\dot{\ell}(t,\beta,\Lambda)$ and $\ddot{\ell}(t,\beta,\Lambda)$ are bounded, Lipschitz continuous of $\Lambda_t(\beta,\Lambda)$, it is not difficult to verify that $\mathcal{F}_1$ is $P_0$-Donsker and $\mathcal{F}_2$ is $P_0$-Glivenko-Cantelli for this example, with suitably integrable respective envelope functions. Thus all of the conditions of Theorem 19.5 are satisfied.

### *19.3.2 The Profile Sampler*

The material for this section comes largely from Lee, Kosorok and Fine (2005) who proposed inference based on sampling from a posterior distribution based on the profile likelihood. The quadratic expansion of the

profile likelihood given in the previous section permits the construction of confidence sets for $\theta$ by inverting the log-likelihood ratio. Translating this elegant theory into practice can be computationally challenging. Even if the log profile likelihood ratio can be successfully inverted for a multivariate parameter, this inversion does not enable the construction of confidence intervals for each one-dimensional subcomponent separately, as is standard practice in data analysis. For such confidence intervals, it would be necessary to further profile over all remaining components in $\theta$. A related problem for which inverting the log likelihood is not adequate is the construction of rectangular confidence regions for $\theta$, such as minimum volume confidence rectangles (Di Bucchiano, Einmahl and Mushkudiani, 2001) or rescaled marginal confidence intervals. For many practitioners, rectangular regions are preferable to ellipsoids, for ease of interpretation.

In principle, having an estimator of $\theta$ and its variance simplifies these inferences considerably. However, the computation of these quantities using the semiparametric likelihood poses stiff challenges relative to those encountered with parametric models, as has been illustrated in several places in this book. Finding the maximizer of the profile likelihood is done implicitly and typically involves numerical approximations. When the nuisance parameter is not $\sqrt{n}$ estimable, nonparametric functional estimation of $\eta$ for fixed $\theta$ may be required, which depends heavily on the proper choice of smoothing parameters. Even when $\eta$ is estimable at the parametric rate, and without smoothing, $\tilde{I}_0$ does not ordinarily have a closed form. When it does have a closed form, it may include linear operators which are difficult to estimate well, and inverting the estimated linear operators may not be straightforward. The validity of these variance estimators must be established on a case-by-case basis.

The bootstrap is a possible solution to some of these problems, and we will discuss this later, briefly, in this chapter and in more detail in Chapter 21. We will show that theoretical justification for the bootstrap is possible but quite challenging for semiparametric models where the nuisance parameter is not $\sqrt{n}$ consistent. Even when the bootstrap can be shown to be valid, the computational burden is quite substantial, since maximization over both $\theta$ and $\eta$ is needed for each bootstrap sample. A different approach to variance estimation is possible via Corollary 19.4 above which verifies that the curvature of the profile likelihood near $\hat{\theta}_n$ is asymptotically equal to $\tilde{I}_0$. In practice, one can perform second order numerical differentiation by evaluating the profile likelihood on a hyperrectangular grid of $3^k$ equidistant points centered at $\hat{\theta}_n$, taking the appropriate differences, and then dividing by $4h^2$, where $p$ is the dimension of $\theta$ and $h$ is the spacing between grid points. While the properties of $h$ for the asymptotic validity of this approach are well known, there are no clear cut rules on choosing the grid spacing in a given data set. Thus, it would seem difficult to automate this technique for practical usage.

Prior to the Lee, Kosorok and Fine (2005) paper, there does not appear to exist in the statistical literature a general theoretically justified and automatic method for approximating $\tilde{I}_0$. Lee, Kosorok and Fine propose an application of Markov chain Monte Carlo to the semiparametric profile likelihood. The method involves generating a Markov chain $\{\theta^{(1)}, \theta^{(2)}, \dots\}$ with stationary density proportional to $p_{\theta,n}(\theta) \equiv \exp(pL_n(\theta))\, q(\theta)$, where $q(\theta) = Q(d\theta)/(d\theta)$ for some prior measure $Q$. This can be accomplished by using, for example, the Metropolis-Hastings algorithm (Metropolis, et al., 1953; and Hastings, 1970). Begin with an initial value $\theta^{(1)}$ for the chain. For each $k = 2, 3, \dots$, obtain a proposal $\tilde{\theta}^{k+1}$ by random walk from $\theta^{(k)}$. Compute $p_{\tilde{\theta}^{k+1},n}(\tilde{\theta}^{k+1})$, and decide whether to accept $\tilde{\theta}^{k+1}$ by evaluating the ratio $p_{\tilde{\theta}^{k+1},n}(\tilde{\theta}^{k+1})/p_{\theta^{(k)},n}(\theta^{(k)})$ and applying an acceptance rule. After generating a sufficiently long chain, one may compute the mean of the chain to estimate the maximizer of $pL_n(\theta)$ and the variance of the chain to estimate $\tilde{I}_0^{-1}$. The output from the Markov chain can also be directly used to construct various confidence sets, including minimum volume confidence rectangles. Whether or not a Markov chain is used to sample from the "posterior" proportional to $\exp(pL_n(\theta))\, q(\theta)$, the procedure based on sampling from this posterior is referred to as the *profile sampler*.

Part of the computational simplicity of this procedure is that $pL_n(\theta)$ does not need to be maximized, it only needs to be evaluated. As mentioned earlier in this chapter, the profile likelihood is generally fairly easy to compute as a consequence of algorithms such as the stationary point algorithm for maximizing over the nuisance parameter. On the other hand, sometimes the profile likelihood can be very hard to compute. When this is the case, numerical differentiation via Corollary 19.4 may be advantageous since it requires fewer evaluations of the profile likelihood. However, numerical evidence in Section 4.2 of Lee, Kosorok and Fine (2005) seems to indicate that, at least for moderately small samples, numerical differentiation does not perform as well in general as the profile sampler. This observation is supported by theoretical work on the profile sampler by Cheng and Kosorok (2007a, 2007b) who show that the profile sampler yields frequentist inference that is second-order accurate. Thus the profile sampler may be beneficial even when the profile likelihood is hard to compute.

The procedure's validity is established in Theorem 19.6 below which extends Theorem 19.5 in a manner that enables the quadratic expansion of the log-likelihood around $\hat{\theta}_n$ to be valid in a fixed, bounded set, rather than only in a shrinking neighborhood. The conclusion of these arguments is that the "posterior" distribution of the profile likelihood with respect to a prior on $\theta$ is asymptotically equivalent to the distribution of $\hat{\theta}_n$. In order to do this, the new theorem will require an additional assumption on the profile likelihood. Define $\Delta_n(\theta) \equiv n^{-1}(pL_n(\theta) - pL_n(\hat{\theta}_n))$. Here is the theorem:

THEOREM 19.6 *Assume $\Theta$ is compact, $\tilde{I}_0$ is positive definite, $Q(\Theta) < \infty$, $q(\theta_0) > 0$, and $q$ is continuous at $\theta_0$. Assume also that $\hat{\theta}_n$ is efficient and that (19.10) holds for $\check{\theta}_n = \hat{\theta}_n$ and for any possibly random sequence $\tilde{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$. Assume moreover that for every random sequence $\tilde{\theta}_n \in \Theta$*

$$\Delta_n(\tilde{\theta}_n) = o_{P_0}(1) \quad \textit{implies that} \quad \tilde{\theta}_n = \theta_0 + o_{P_0}(1). \tag{19.17}$$

*Then, for every measurable function $g : \mathbb{R}^k \mapsto \mathbb{R}$ satisfying*

$$\limsup_{k\to\infty} k^{-2} \log \left( \sup_{u\in\mathbb{R}^k : \|u\| \le k} |g(u)| \right) \le 0, \tag{19.18}$$

*we have*

$$\begin{aligned} &\frac{\int_\Theta g\left(\sqrt{n}(\theta - \hat{\theta}_n)\right) p_{\theta,n}(\theta) d\theta}{\int_\Theta p_{\theta,n} d\theta} \\ &\quad = \int_{\mathbb{R}^k} g(u)(2\pi)^{-k/2} |\tilde{I}_0|^{1/2} \exp\left[-\frac{u'\tilde{I}_0 u}{2}\right] du + o_{P_0}(1). \end{aligned} \tag{19.19}$$

The proof is given in Section 19.4 below. Note that when $g(u) = O(1 + \|u\|)^d$, for any $d < \infty$, Condition (19.18) is readily satisfied. This means that indicators of measurable sets and the first two moments of $\sqrt{n}(T - \hat{\theta}_n)$, where $T$ has the posterior density proportional to $t \mapsto p_{t,n}(t)$, are consistent for the corresponding probabilities and moments of the limiting Gaussian distribution. Specifically, $\mathrm{E}(T) = \hat{\theta}_n + o_{P_0}(n^{-1/2})$ and $n\mathrm{var}(T) = \tilde{I}_0^{-1} + o_{P_0}(1)$. Thus we can calculate all the quantities needed for inference on $\theta$ without having to actually maximize the profile likelihood directly or compute derivatives.

Note that the interesting Condition (19.17) is not implied by the other conditions and is not implied by the identifiability of the Kulback-Leibler information from the full likelihood. Nevertheless, if it can be shown that $\Delta_n(\theta)$ converges uniformly over $\Theta$ to the profiled Kulback-Leibler information $\Delta_0(\theta)$, then identifiability of the Kulback-Leibler information for $L_n(\theta, \eta)$ is sufficient. This approach works for the Cox model for right-censored data, as we will see below. However, in two out of the three examples considered in the Lee, Kosorok and Fine (2005) paper, and probably for a large portion of models generally, it appears that the strategy based on $\Delta_0(\theta)$ is usually not fruitful, and it seems to be easier to establish (19.17) directly. The Condition (19.17) is needed because the integration in (19.19) is over all of $\Theta$, and thus it is important to guarantee that there are no other distinct modes besides $\hat{\theta}_n$ in the limiting posterior. Condition (19.10) is not sufficient for this since it only applies to shrinking neighborhoods of $\theta_0$ and not to all of $\Theta$ as required.

The examples and simulation studies in Lee, Kosorok and Fine demonstrate that the profile sampler works very well and is in general computationally efficient. The Metropolis algorithm applied to $p_{\theta,n}(\theta)$ with a Lebesgue prior measure is usually quite easy to tune and seems to achieve equilibrium quickly. By the ergodic theorem, there exists a sequence of finite chain lengths $\{M_n\} \to \infty$ so that the chain mean $\bar{\theta}_n \equiv M_n^{-1} \sum_{j=1}^{M_n} \theta^{(j)}$ satisfies $\bar{\theta}_n = \hat{\theta}_n + o_{P_0}(n^{-1/2})$; the standardized sample variance $V_n \equiv M_n^{-1} \sum_{j=1}^{M_n} n(\theta^{(i)} - \bar{\theta}_n)(\theta^{(i)} - \bar{\theta}_n)'$ is consistent for $\tilde{I}_0^{-1}$; and the empirical measure $G_n(A) \equiv M_n^{-1} \sum_{j=1}^{M_n} 1\left\{\sqrt{n}(\theta^{(j)} - \bar{\theta}_n) \in A\right\}$, for a bounded convex $A \subset \mathbb{R}^k$, is consistent for the probability that a mean zero Gaussian deviate with variance $\tilde{I}_0^{-1}$ lies in $A$. Hence the output of the chain can be used for inference about $\theta_0$, provided $M_n$ is large enough so that the sampling error from using a finite chain is negligible.

We now verify the additional Assumption (19.17) for the Cox model for right censored data and for the Cox model for current status data. Note that the requirements that $\hat{\theta}_n$ be efficient and that (19.10) hold, for $\breve{\theta}_n = \hat{\theta}_n$ and for any possibly random sequence $\tilde{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$, have already been established for these examples at the end of the previous section. Thus verifying (19.17) will enable the application of Theorem 19.6 to these examples.

*The Cox model for right censored data.*

For this example, we can use the identifiability of the profile Kulback-Leibler information since the profile likelihood does not involve the nuisance parameter. Let $B$ be the compact parameter space for $\beta$, where $\beta_0$ is known to be in the interior of $B$, and assume that $\|Z\|$ is bounded by a constant. We know from our previous discussions of this model that $n^{-1}pL_n(\beta)$ equals, up to a constant that does not depend on $\beta$,

$$H_n(\beta) \equiv \mathbb{P}_n \left[ \int_0^\tau \left( \beta' Z - \log \left[ \mathbb{P}_n Y(s) e^{\beta' Z} \right] \right) dN(s) \right].$$

By arguments which are by now familiar to the reader, it is easy to verify that $\|H_n - H_0\|_B \xrightarrow{\mathrm{P}} 0$, where

$$H_0(\beta) \equiv P_0 \left[ \int_0^\tau \left( \beta' Z - \log P_0 \left[ Y(s) e^{\beta' Z} \right] \right) dN(s) \right].$$

It is also easy to verify that $H_0$ has first derivative

$$U_0(\beta) \equiv P_0 \left[ \int_0^\tau \left( Z - E(s, \beta) \right) dN(s) \right],$$

where $E(s, \beta)$ is as defined in Section 4.2.1, and second derivative $-V(\beta)$, where $V(\beta)$ is defined in (4.5). By the boundedness of $\|Z\|$ and $B$ combined with the other assumptions of the model, it can be shown (see Exercise 19.5.8 below) that there exists a constant $c_0 > 0$ not depending on

$\beta$ such that $V(\beta) \geq c_0 \mathrm{var} Z$, where for $k \times k$ matrices $A$ and $B$, $A \geq B$ means that $c'Ac \geq c'Bc$ for every $c \in \mathbb{R}^k$. Thus $H_0$ is strictly concave and thus has a unique maximum on $B$. It is also easy to verify that $U_0(\beta_0) = 0$ (see Part (b) of Exercise 19.5.8), and thus the unique maximum is located at $\beta = \beta_0$.

Hence $\|\Delta_n(\beta) - \Delta_0(\beta)\|_B \xrightarrow{\mathrm{P}} 0$, where $\Delta_0(\beta) = H_0(\beta) - H_0(\beta_0) \leq 0$ is continuous, with the last inequality being strict whenever $\beta \neq \beta_0$. This immediately yields Condition (19.17) for $\beta$ replacing $\theta$.

*The Cox model for current status data.*

For this example, we verify (19.17) directly. Let $\tilde{\beta}_n$ be some possibly random sequence satisfying $\Delta_n(\tilde{\beta}_n) = o_{P_0}(1)$, where $\beta$ is replacing $\theta$. Fix some $\alpha \in (0,1)$ and note that since $\Delta_n(\tilde{\beta}_n) = o_{P_0}(1)$ and $\Delta_n(\beta_0) \leq 0$ almost surely, we have

$$n^{-1} \sum_{i=1}^{n} \log \left\{ \frac{f(\tilde{\beta}_n, \hat{F}_{\tilde{\beta}_n}; X_i)}{f(\beta_0, F_0; X_i)} \right\} \geq o_{P_0}(1),$$

where $f(\beta, F; X) \equiv \delta \left\{ 1 - \overline{F}(Y)^{\exp(\beta' Z)} \right\} + (1 - \delta) \overline{F}(Y)^{\exp(\beta' Z)}$, $\overline{F} \equiv 1 - F = \exp(-\Lambda)$, and $\hat{F}_\beta \equiv 1 - \exp(-\hat{\Lambda}_\beta)$ is the maximizer of the likelihood over the nuisance parameter for fixed $\beta$. This now implies

$$n^{-1} \sum_{i=1}^{n} \log \left[ 1 + \alpha \left\{ \frac{f(\tilde{\beta}_n, \hat{F}_{\tilde{\beta}_n}; X_i)}{f(\beta_0, F_0; X_i)} - 1 \right\} \right] \geq o_{P_0}(1),$$

because $\alpha \log(x) \leq \log(1 + \alpha\{x - 1\})$ for any $x > 0$. This implies that

$$P_0 \log \left[ 1 + \alpha \left\{ \frac{f(\tilde{\beta}_n, \hat{F}_{\tilde{\beta}_n}; X_i)}{f(\beta_0, F_0; X_i)} - 1 \right\} \right] \quad \geq \quad o_{P_0}(1) \tag{19.20}$$

by Lemma 19.7 below, the proof of which is given in Section 19.4, since $x \mapsto \log(1 + \alpha x)$ is Lipschitz continuous for $x \geq 0$ and $f(\theta_0, F_0; X) \geq c$ almost surely, for some $c > 0$.

Because $x \mapsto \log x$ is strictly concave, we now have that

$$P_0 \log \left[ 1 + \alpha \left\{ \frac{f(\tilde{\beta}_n, \hat{F}_{\tilde{\beta}_n}; X_i)}{f(\beta_0, F_0; X_i)} - 1 \right\} \right] \leq 0.$$

This combined with (19.20) implies that

$$P_0 \log \left[ 1 + \alpha \left\{ \frac{f(\tilde{\beta}_n, \hat{F}_{\tilde{\beta}_n}; X_i)}{f(\beta_0, F_0; X_i)} - 1 \right\} \right] = o_{P_0}(1).$$

This forces the result

$$P_0 \left| \overline{F}_{\tilde{\beta}_n}(Y)^{\exp(\tilde{\beta}_n' Z)} - \overline{F}_0(Y)^{\exp(\beta_0' Z)} \right| = o_{P_0}(1)$$

by the strict concavity of $x \mapsto \log x$. This, in turn, implies that

$$P_0 \left[ \left\{ (\tilde{\beta}_n - \beta_0)'(Z - E[Z|Y]) - c_n(Y) \right\}^2 \middle| Y \right] = o_{P_0}(1),$$

for almost surely all $Y$, where $c_n(Y)$ is uncorrelated with $Z - \mathrm{E}[Z|Y]$. Hence $\tilde{\beta}_n = \theta_0 + o_{P_0}(1)$, and Condition (19.17) now follows.

LEMMA 19.7 *The class* $\mathcal{F} \equiv \{f(\beta, F; X) : \beta \in B, F \in \mathcal{M}\}$, *where* $\mathcal{M}$ *is the class of distribution functions on* $[0, \tau]$, *is* $P_0$*-Donsker.*

### *19.3.3 The Penalized Profile Sampler*

In many semiparametric models involving a smooth nuisance parameter, it is often convenient and beneficial to perform estimation using penalization. One motivation for this is that, in the absence of any restrictions on the form of the function $\eta$, maximum likelihood estimation for some semiparametric models leads to over-fitting. We have discussed this issue in the context of the partly linear logistic regression model (see Chapter 1 and Sections 4.5, 15.1 and, in this chapter, Section 19.2.4). Under certain reasonable regularity conditions, penalized semiparametric log-likelihood estimation can yield fully efficient estimates for $\theta$, as we demonstrated in Section 4.5.

The purpose of this section is to briefly discuss a modification of the profile sampler that works with profiled penalized likelihoods, called the *penalized profile sampler*. Another method of inference that works in this context is a weighted bootstrap which will be discussed in Chapter 21 and is applicable in general to semiparametric M-estimation. Interestingly, it is unclear whether the usual nonparametric bootstrap will work at all in this this context.

We assume in this section that the nuisance parameter $\eta$ is a function in Sobolev class of functions supported on some compact set $\mathcal{U}$ on the real line, whose $d$-th derivative exists and is absolutely continuous with $J(\eta) < \infty$, where

$$J^2(\eta) = \int_{\mathcal{U}} (\eta^{(d)}(u))^2 du.$$

Here $d$ is a fixed, positive integer and $\eta^{(j)}$ is the $j$-th derivative of $\eta$ with respect to $u$. Obviously $J^2(\eta)$ is some measurement of the complexity of $\eta$. We denote $\mathcal{H}$ to be the Sobolev function class with degree $d$ on $\mathcal{U}$.

The penalized log-likelihood in this context is:

$$\tilde{L}_n(\theta, \eta) = n^{-1} L_n(\theta, \eta) - \lambda_n^2 J^2(\eta),$$

where $L_n(\theta, \eta)$ is the log-likelihood as used in previous sections in this chapter, and $\lambda_n$ is a smoothing parameter, possibly dependent on the data. This is the approach we have been using for the partly linear logistic regression model. We make the same assumptions about the smoothing parameter in this general context that we made for the partly linear logistic regression model, namely:

$$\lambda_n = o_{P_0}(n^{-1/4}) \text{ and } \lambda_n^{-1} = O_{P_0}(n^{d/(2d+1)}). \tag{19.21}$$

One way to ensure (19.21) in practice is simply to set $\lambda_n = n^{-d/(2d+1)}$. Or we can just choose $\lambda_n = n^{-1/3}$ which is independent of $d$. The log-profile penalized likelihood is defined as $p\tilde{L}_n(\theta) = \tilde{L}_n(\theta, \tilde{\eta}_\theta)$, where $\tilde{\eta}_\theta$ is $\text{argmax}_{\eta \in \mathcal{H}} \tilde{L}_n(\theta, \eta)$ for fixed $\theta$ and $\lambda_n$. The *penalized profile sampler* is actually just the procedure of sampling from the posterior distribution of $p\tilde{L}_n(\theta)$ by assigning a prior on $\theta$ (Cheng and Kosorok, 2007c).

Cheng and Kosorok (2007c) analyze the penalized profile sampler and obtain the following conclusions under reasonable regularity conditions:

- **Distribution approximation**: The posterior distribution proportional to $\exp\left(p\tilde{L}_n(\theta)\right) q(\theta)$, where $q$ is a suitably smooth prior density with $q(\theta_0) > 0$, can be approximated by a normal distribution with mean the maximum penalized likelihood estimator of $\theta$ and variance the inverse of the efficient information matrix, with a level of accuracy similar to that obtained for the profile sampler (see Theorem 19.6).

- **Moment approximation**: The maximum penalized likelihood estimator of $\theta$ can be approximated by the posterior mean with error $O_{P_0}(\lambda_n^2)$. The inverse of the efficient information matrix can be approximated by the posterior variance with error $O_{P_0}(n^{1/2}\lambda_n^2)$.

- **Confidence interval approximation**: An exact frequentist confidence interval of Wald's type for $\theta$ can be estimated by the credible set obtained from the posterior distribution with error $O_{P_0}(\lambda_n^2)$.

The last item above is perhaps the most important: a confidence region based on the posterior distribution will be $O_{P_0}(\lambda_n^2)$ close to an exact frequentist confidence region. This means, for example, that when $d = 2$, and if $\lambda_n = n^{-2/5}$, the size of the error would be $O_{P_0}(n^{-4/5})$. Note that this is second order accurate since the first order coverage accuracy of a parametric confidence band is $o_{P_0}(n^{-1/2})$.

Cheng and Kosorok (2007c) also verify that the partly linear logistic regression model satisfies the needed regularity conditions for the penalized profile sampler with the approximately least-favorable submodel described in Section 19.2.4.

### *19.3.4 Other Methods*

In this section, we briefly discuss a few alternatives to quadratic expansion and the profile sampler. We first discuss the bootstrap and some related procedures. We then present a computationally simpler alternative, the "block jackknife," and then briefly discuss a fully Bayesian alternative. Note that this list is not exhaustive, and there are a number of useful methods that we are omitting.

The nonparametric bootstrap is useful for models where all parameters are $\sqrt{n}$-consistent, but it is unclear how well it works when the nuisance parameter is not $\sqrt{n}$-consistent. The weighted bootstrap appears to be more applicable in general, and we will discuss this method in greater detail in Chapter 21 in the context of general semiparametric M-estimators, including non-likelihood based approaches. Two other important alternatives are the $m$ within $n$ bootstrap (see Bickel, Götze and van Zwet, 1997) and subsampling (see Politis and Romano, 1994). The idea for both of these is to compute the estimate based on a bootstrap sample of size $m < n$. For the $m$ within $n$ bootstrap, the sample is taken with replacement, while for subsampling, the sample is taken without replacement. The properties of both procedures are quite similar when $m/n \to 0$ and both $m$ and $n$ go to infinity. The main result of Politis and Ramano is that the subsampling boots the true limiting distribution whenever the true limiting distribution exists and is continuous. Since $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is known to have a continuous limiting distribution $\mathcal{L}$, Theorem 2.1 of Politis and Romano (1994) yields that the $m$ out of $n$ subsampling bootstrap converges—conditionally on the data—to the same distribution $\mathcal{L}$, provided $m/n \to 0$ and $m \to \infty$ as $n \to \infty$.

Because of the requirement that $m \to \infty$ as $n \to \infty$, the subsampling bootstrap potentially involves many calculations of the estimator. Fortunately, the asymptotic linearity of $\hat{\theta}_n$ can be used to formulate a computationally simpler alternative, the "block jackknife," as described in Ma and Kosorok (2005a). Let $\tilde{\theta}_n$ be any asymptotically linear estimator of a parameter $\theta_0 \in \mathbb{R}^k$, based on an i.i.d. sample $X_1, \ldots, X_n$, having square-integrable influence function $\phi$ for which $\mathrm{E}[\phi\phi^T]$ is nonsingular.

Let $m$ be a fixed integer $> k$, and, for each $n \geq m$, define $q_{m,n}$ to be the largest integer satisfying $mq_{m,n} \leq n$. Also define $N_{m,n} \equiv mq_{m,n}$. For the data $X_1, \ldots, X_n$, compute the estimator $\tilde{\theta}_n$ and randomly sample $N_{m,n}$ out of the $n$ observations without replacement, to obtain $X_1^*, \ldots, X_{N_{m,n}}^*$. Note that we are using the notation $\tilde{\theta}_n$ rather than $\hat{\theta}_n$ to remind ourselves that this estimator is a general, asymptotically linear estimator and not necessarily a maximum likelihood estimator. For $j = 1, \ldots, m$, let $\tilde{\theta}_{n,j}^*$ be the estimate of $\theta$ based on the observations $X_1^*, \ldots, X_{N_{m,n}}^*$ after omitting $X_j^*, X_{m+j}^*, X_{2m+j}^*, \ldots, X_{(q_{m,n}-1)m+j}^*$. Compute $\bar{\theta}_n^* \equiv m^{-1} \sum_{j=1}^m \tilde{\theta}_{n,j}^*$ and $S_n^* \equiv (m-1)q_{m,n} \sum_{j=1}^m \left(\tilde{\theta}_{n,j}^* - \bar{\theta}_n^*\right)\left(\tilde{\theta}_{n,j}^* - \bar{\theta}_n^*\right)^T$. The following lemma

provides a method of obtaining asymptotically valid confidence ellipses for $\theta_0$:

LEMMA 19.8 *Let $\tilde{\theta}_n$ be an estimator of $\theta_0 \in \mathbb{R}^k$, based on an i.i.d. sample $X_1, \ldots, X_n$, which satisfies $n^{1/2}(\tilde{\theta}_n - \theta_0) = \sqrt{n}\mathbb{P}_n\phi + o_p(1)$, where $\mathrm{E}[\phi\phi^T]$ is nonsingular. Then $n(\tilde{\theta}_n - \theta_0)^T [S_n^*]^{-1} (\tilde{\theta}_n - \theta_0)$ converges weakly to $k(m-1)F_{k,m-k}/(m-k)$, where $F_{r,s}$ has an F distribution with respective degrees of freedom $r$ and $s$.*

**Proof.** Fix $m > k$. Without loss of generality, we can assume by the i.i.d. structure that $X_i^* = X_i$ for $i = 1, \ldots, N_{m,n}$. Let $\epsilon_{0,n} \equiv \sqrt{n}(\tilde{\beta}_n - \beta_0) - n^{-1/2} \sum_{i=1}^n \phi_i$ and

$$\epsilon_{j,n} \equiv (N_{m,n} - m)^{1/2}(\tilde{\beta}_{n,j}^* - \beta_0) - (N_{m,n} - m)^{-1/2} \sum_{i \in K_{j,n}} \phi_i,$$

where $K_{j,n} \equiv \{1, \ldots, n\} - \{j, m+j, 2m+j, \ldots, (q_{m,n}-1)m + j\}$, for $j = 1, \ldots, m$; and note that $\max_{0 \le j \le m} |\epsilon_{j,n}| = o_{P_0}(1)$ by asymptotic linearity. Now let $Z_{j,n}^* \equiv q_{m,n}^{-1/2} \sum_{i=1}^{q_{m,n}} \phi_{(i-1)m+j}$, for $j = 1 \ldots m$, and define $\bar{Z}_n^* \equiv m^{-1} \sum_{j=1}^m Z_{j,n}^*$. Thus $S_n^* = (m-1)^{-1} \sum_{j=1}^m (Z_{j,n}^* - \bar{Z}_n^*)(Z_{j,n}^* - \bar{Z}_n^*)^T + o_p(1)$. Hence $S_n^*$ and $\sqrt{n}(\tilde{\beta}_n - \beta_0)$ are jointly asymptotically equivalent to $S_m \equiv (m-1)^{-1} \sum_{j=1}^m (Z_j - \bar{Z}_m)(Z_j - \bar{Z}_m)^T$ and $Z_0$, respectively, where $\bar{Z}_m \equiv m^{-1} \sum_{j=1}^m Z_j$ and $Z_0, \ldots, Z_m$ are i.i.d. mean zero Gaussian deviates with variance $\mathrm{E}[\phi\phi^T]$. Now the results follow by standard normal theory (see Appendix V of Scheffé, 1959).□

The fact that $m$ remains fixed as $n \to \infty$ in the above approach results in a potentially significant computational savings over subsampling which requires $m$ to grow increasingly large as $n \to \infty$. A potential challenge is in choosing $m$ for a given data set. The larger $m$ is, the larger the denominator degrees of freedom in $F_{d,m-d}$ and the tighter the confidence ellipsoid. On the other hand, $m$ cannot be so large that the required asymptotic linearity does not hold simultaneously for all jackknife components. The need to choose $m$ makes this approach somewhat less automatic than the profile sampler. Nevertheless, that fact that this "block jackknife" procedure requires fewer assumptions than the profile sampler makes it a potentially useful alternative.

Of course, it is always simplest to estimate the variance directly whenever this is possible. Unfortunately, this is rarely the case except for extremely simple semiparametric models such as the Cox model for right censored data. Inference on $\theta$ can also be based on the marginal posterior of $\theta$ from the full likelihood with respect to a joint prior on $(\theta, \eta)$. Shen (2002) has shown that this approach yields valid inferences for $\hat{\theta}_n$ when $\theta$ is estimable at the parametric rate. The profile sampler, however, appears to be a better choice most of the time since it greatly simplifies the theory and computations through the avoidance of specifying a prior for $\eta$. In this light, the

profile sampler may be useful as an approximately Bayesian alternative to a fully Bayesian procedure when $\eta$ is strictly a nuisance parameter.

## 19.4 Proofs

**Proof of Theorem 19.6.** Let

$$\delta_n^2 \equiv n^{-1} \vee \sup_{k \geq n} k^{-1} \log \left\{ \sup_{\theta_1, \theta_2 \in \Theta} \left| g\left(\sqrt{k}[\theta_1 - \theta_2]\right) \right| \right\},$$

where $x \vee y$ denotes the maximum of $x$ and $y$, and note that by Condition (19.18) this defines a positive sequence $\{\delta_n\} \downarrow 0$ such that $n\delta_n \to \infty$. Hence also $\limsup_{n\to\infty} r_n \leq 0$, where

$$r_n = \sup_{\theta_1, \theta_2 \in \Theta} \frac{\log |\sqrt{n} g\{\sqrt{n}(\theta_1 - \theta_2)\}|}{n\delta_n}.$$

Fix $\epsilon > 0$, and let

$$\Delta_n^\epsilon \equiv \sup_{\theta \in \Theta : \|\theta - \hat{\theta}_n\| > \epsilon} \Delta_n(\theta).$$

Now note that $p_{\theta,n}(\theta)$ is proportional to $\exp\{n\Delta_n(\theta)\}q(\theta)$ and that

$$\begin{aligned} &\int_{\theta \in \Theta : \|\theta - \hat{\theta}\| > \epsilon} \sqrt{n} \left| g\{\sqrt{n}(\theta - \hat{\theta}_n)\} \right| \exp\{n\Delta_n(\theta)\} q(\theta) d\theta \\ &\quad \leq \quad 1\{\Delta_n^\epsilon < -\delta_n\} \int_\Theta q(\theta) d\theta \exp\{-n\delta_n[1 - r_n]\} + o_{P_0}(1) \\ &\quad \to \quad 0, \end{aligned}$$

in probability, by Lemma 19.9 below. This now implies that there exists a positive decreasing sequence $\{\epsilon_n\} \downarrow 0$ such that $\sqrt{n}\epsilon_n \to \infty$ and

$$\int_{\theta \in \Theta : \|\theta - \hat{\theta}_n\| > \epsilon_n} \sqrt{n} |g\{\sqrt{n}(\theta - \hat{\theta}_n)\}| \exp\{n\Delta_n(\theta)\} q(\theta) d\theta \to 0,$$

in probability.

Now,

$$
\begin{aligned}
&\int_{\theta\in\Theta:\|\theta-\hat\theta_n\|\le\epsilon_n} \sqrt{n}|g\{\sqrt{n}(\theta-\hat\theta_n)\}| \exp\left\{-\frac{n}{2}(\theta-\hat\theta_n)'\tilde I_0(\theta-\hat\theta_n)\right\} \\
&\qquad \times \left|\exp\left\{n\Delta_n(\theta)+\frac{n}{2}(\theta-\hat\theta_n)'\tilde I_0(\theta-\hat\theta_n)\right\}-1\right| q(\theta)d\theta \\
&\le \int_{\theta\in\Theta:\|\theta-\hat\theta_n\|\le\epsilon_n} \sqrt{n}|g\{\sqrt{n}(\theta-\hat\theta_n)\}| \\
&\qquad\qquad \times \exp\left\{-\frac{n}{4}(\theta-\hat\theta_n)'\tilde I_0(\theta-\hat\theta_n)\right\} q(\theta)d\theta \\
&\qquad \times \sup_{\theta\in\Theta:\|\theta-\hat\theta_n\|\le\epsilon_n} \left[\exp\left\{-\frac{n}{4}(\theta-\hat\theta_n)'\tilde I_0(\theta-\hat\theta_n)\right\}\right. \\
&\qquad\qquad \left. \times \left|\exp\left\{n\Delta_n(\theta)+\frac{n}{2}(\theta-\hat\theta_n)'\tilde I_0(\theta-\hat\theta_n)\right\}-1\right|\right] \\
&= A_n \times \sup_{\theta\in\Theta:\|\theta-\hat\theta_n\|\le\epsilon_n} B_n(\theta),
\end{aligned}
$$

where $A_n = O_{P_0}(1)$ by Condition (19.18). However, for any random sequence $\tilde\theta_n \in \Theta$ such that $\|\tilde\theta_n - \hat\theta_n\| \le \epsilon_n$ for all $n \ge 1$, and any fixed $M < \infty$, we have by Condition (19.10) that

$$
\begin{aligned}
B_n(\tilde\theta_n) &\le \exp\left\{-\frac{n}{4}(\tilde\theta_n-\hat\theta_n)'\tilde I_0(\tilde\theta_n-\hat\theta_n)\right\} \\
&\qquad \times \left|\exp\left\{o_{P_0}\left(1+\sqrt{n}\|\tilde\theta_n-\theta_0\|\right)^2\right\}-1\right|,
\end{aligned} \tag{19.22}
$$

where

$$
\begin{aligned}
o_{P_0}\left(1+\sqrt{n}\|\tilde\theta_n-\theta_0\|\right)^2 &\le o_{P_0}(1)\left[\left(1+\sqrt{n}\|\tilde\theta_n-\hat\theta_n\|\right)^2 + n\|\hat\theta_n-\theta_0\|^2\right] \\
&= o_{P_0}(1)\left(1+\sqrt{n}\|\tilde\theta_n-\hat\theta_n\|\right)^2,
\end{aligned}
$$

since $\sqrt{n}\|\hat\theta_n-\theta_0\| = O_{P_0}(1)$. Combining this with (19.22) and the fact that $|e^x-1| \le e^{|x|}$ for all $x \in \mathbb{R}$, we obtain that

$$
\begin{aligned}
B_n(\tilde\theta_n) &\le \exp\left\{-\frac{n}{4}(\tilde\theta_n-\hat\theta_n)'\tilde I_0(\tilde\theta_n-\hat\theta_n)\right\} \\
&\qquad\qquad \times |\exp\{o_{P_0}(1)\}-1|\, 1\left\{\sqrt{n}\|\tilde\theta_n-\hat\theta_n\| \le M\right\} \\
&\qquad + \exp\left\{-\frac{n}{4}(\tilde\theta_n-\hat\theta_n)'\tilde I_0(\tilde\theta_n-\hat\theta_n)\right. \\
&\qquad\qquad\qquad \left. + o_{P_0}(1)\left(1+\sqrt{n}\|\tilde\theta_n-\hat\theta_n\|\right)^2\right\} \\
&\qquad \times 1\left\{\sqrt{n}\|\tilde\theta_n-\hat\theta_n\| > M\right\} \\
&\le o_{P_0}(1) + \exp\left\{-\frac{M^2}{4}\lambda_1(1-o_{P_0}(1))\right\} \\
&\to \exp\left\{-\frac{M^2}{4}\lambda_1\right\},
\end{aligned}
$$

in probability, where $\lambda_1$ is the smallest eigenvalue of $\tilde{I}_0$ and is $> 0$ by positive definiteness. Hence $B_n(\tilde{\theta}_n) = o_{P_0}(1)$ since $M$ was arbitrary. Thus $\sup_{\theta\in\Theta:\|\theta-\hat{\theta}_n\|\le\epsilon_n} B_n(\theta) = o_{P_0}(1)$.

By reapplication of Condition (19.9),

$$\begin{aligned}
&\int_{\theta\in\Theta:\|\theta-\hat{\theta}_n\|\le\epsilon_n} \sqrt{n}g\{\sqrt{n}(\theta-\hat{\theta}_n)\}\exp\left\{-\frac{n}{2}(\theta-\hat{\theta}_n)'\tilde{I}_0(\theta-\hat{\theta}_n)\right\}q(\theta)d\theta \\
&= \int_{u\in\sqrt{n}(\Theta-\hat{\theta}_n):\|u\|\le\sqrt{n}\epsilon_n} g(u)\exp\left\{-\frac{u'\tilde{I}_0u}{2}\right\}q\left(\hat{\theta}_n+\frac{u}{\sqrt{n}}\right)du \\
&\to \int_{\mathbb{R}^k} g(u)\exp\left\{-\frac{u'\tilde{I}_0u}{2}\right\}du\times q(\theta_0),
\end{aligned}$$

in probability. (Here, we have used the notation $\sqrt{n}(\Theta-\hat{\theta}_n)$ to denote the set $\{\sqrt{n}(\theta-\hat{\theta}_n):\theta\in\Theta\}$.) The same conclusion holds true for $g=1$, and thus

$$\int_\Theta \sqrt{n}\exp\{n\Delta_n(\theta)\}q(\theta)d\theta = \int_{\mathbb{R}^k}\exp\left\{-\frac{u'\tilde{I}_0u}{2}\right\}du\times q(\theta_0)+o_{P_0}(1).$$

Hence the desired result follows.□

LEMMA 19.9 *Assume Condition (19.17). For every $\epsilon > 0$ and every positive decreasing sequence $\{\delta_n\}\downarrow 0$,*

$$\lim_{n\to\infty}\mathrm{P}\left\{\Delta_n^\epsilon \ge -\delta_n\right\} = 0.$$

**Proof.** Suppose that the lemma is not true. Then there exists an $\epsilon > 0$, a positive decreasing sequence $\{\delta_n^*\}\downarrow 0$, a random sequence $\tilde{\theta}_n\in\Theta$ : $\|\tilde{\theta}_n-\hat{\theta}_n\| > \epsilon$, and a subsequence $\{n_k\}$ such that

$$\lim_{k\to\infty}\mathrm{P}\left\{\Delta_{n_k}(\tilde{\theta}_{n_k}) \ge -\delta_{n_k}^*\right\} = \rho > 0.$$

Define $H_n = \left\{\Delta_n(\tilde{\theta}_n)\ge -\delta_n^*\right\}$, and let $\theta_n^* = 1\{H_n\}\tilde{\theta}_n + 1^cH_n\hat{\theta}_n$, where $1^cA$ is the indicator of the complement of $A$. Now, note that $\Delta_n(\theta_n^*) = \Delta_n(\hat{\theta}_n) = 0$ on $H_n^c$, where $A^c$ is the complement of $A$, and that $\Delta_n(\theta_n^*) = \Delta_n(\tilde{\theta}_n)\ge -\delta_n^*$ on $H_n$. Now $\Delta_n(\theta_n^*)\to 0$ in probability, which implies by Condition (19.17) that $\theta_n^*\to\theta_0$ in probability. Now, on the set $H_{n_k}$, $\theta_{n_k}^* = \tilde{\theta}_{n_k}$ and therefore by construction, $\tilde{\theta}_{n_k}$ is separated from $\hat{\theta}_n$ by at least $\epsilon$. Since $\hat{\theta}_n$ converges to $\theta_0$, we have that eventually, on the set $H_{n_k}$, $\theta_{n_k}^*$ $(=\tilde{\theta}_{n_k})$ is separated from $\theta_0$ by at least $\epsilon/2$. However, the probability of the set $H_{n_k}$ converges to $\rho > 0$, which implies that with probability larger than $\rho/2$, $\theta_{n_k}^*$ is eventually separated from $\theta_0$ by at least $\epsilon/2$. This

contradicts the fact that $\theta^*_{n_k} - \theta_0 \to 0$ in probability, and the desired result follows.□

**Proof of Lemma 19.7.** First note that for all $a, b \in [1, \infty)$,

$$\sup_{u \in [0,1]} |u^a - u^b| \quad \leq \quad c + \log(1/c)|a - b|, \tag{19.23}$$

for any $c > 0$. Minimizing the right-hand-side of (19.23) over $c$, we obtain that the left-hand-side of (19.23) is bounded above by

$$|a - b| \left[ 1 + \log \left( \frac{1}{(|a - b| \wedge 1)} \right) \right]$$

which, in turn, implies that $\sup_{u \in [0,a]} |u^a - u^b| \leq 2|a - b|^{1/2}$ for all $a, b \leq 1$. Here, $x \wedge y$ denotes the minimum of $x$ and $y$. Note also that for any $a \geq 1$ and any $u, v \in [0, 1]$, $|u^a - v^a| \leq a|u - v|$. Let $M \in (1, \infty)$ satisfy $1/M \leq e^{\beta' Z} \leq M$ for all $\beta \in B$ and all possible $Z$. Then, for any $F_1, F_2 \in \mathcal{M}$ and any $\beta_1, \beta_2 \in B$,

$$\begin{aligned} &\left| \overline{F}_1(Y)^{M \exp(\beta_1' Z)} - \overline{F}_2(Y)^{M \exp(\beta_2' Z)} \right| \\ &\qquad \leq \quad M^2 |\overline{F}_1(Y) - \overline{F}_2(Y)| + k|\beta_1 - \beta_2|^{1/2}, \end{aligned} \tag{19.24}$$

for some fixed $k \in (0, \infty)$, since $Z$ lies in a compact set. Since the sets $\{\overline{F} : F \in \mathcal{M}\}$ and $\{\overline{F}^{1/M} : F \in \mathcal{M}\}$ are equivalent, and since the bracketing entropy of $\mathcal{M}$, for brackets of size $\epsilon$, is $O(\epsilon^{-1})$ by Theorem 9.24, (19.24) implies that the bracketing entropy of $\mathcal{F}$ is $O(\epsilon^{-1})$, and the desired result follows.□

## 19.5 Exercises

**19.5.1.** Verify explicitly that $G_n = o_P(1)$, where $G_n$ is defined towards the end of the proof of Theorem 3.1.

**19.5.2.** Verify (19.4).

**19.5.3.** Using the techniques given in the proof of Theorem 15.9, show that $\sigma^{22}_\theta : \mathcal{H}^2_\infty \mapsto \mathcal{H}^2_\infty$ is continuously invertible and onto.

**19.5.4.** In the context of Section 19.2.2, show that the choice $h(Y) = h_0(Y)$, where $h_0$ is as defined in (19.7), is bounded and minimizes $P_{\beta,\Lambda} \|\ell_{\beta,\Lambda} - A_{\beta,\Lambda} h\|^2$ over all square-integrable choices of $h(Y)$.

**19.5.5.** Adapting the arguments leading up to (19.9), prove Corollary 19.1.

**19.5.6.** Use (19.8) and Corollary 19.1 to prove Corollary 19.2.

**19.5.7.** Verify that $\mathcal{F}_1$ is $P_0$-Donsker and that $\mathcal{F}_2$ is $P_0$-Glivenko-Cantelli for the Cox model for right censored data, where $\mathcal{F}_1$ and $\mathcal{F}_2$ are as defined in Theorem 19.5. See the discussion of this model at the end of Section 19.3.1.

**19.5.8.** In the context of the paragraphs on the Cox model for right censored data at the end of Section 19.3.2, show that the following are true:

(a) There exists a constant $c_0 > 0$ not depending on $\beta$ such that $V(\beta) \geq c_0 \text{var} Z$. Hint: It may be helpful to recall that $Y(t) \geq Y(\tau)$ for all $0 \leq t \leq \tau$.

(b) $U_0(\beta_0) = 0$.

## 19.6 Notes

The proof of Theorem 3.1 in Section 19.1 follows closely the proof of Theorem 25.54 in van der Vaart (1998). Corollaries 19.3 and 19.4 are modifications of Corollaries 2 and 3 of Murphy and van der Vaart (2000), while Theorem 19.5 is Murphy and van der Vaart's Theorem 1. Theorem 19.6 is a modification of Theorem 1 of Lee, Kosorok and Fine (2005), while Section 19.3.2 on the Cox model for current status data comes from Section 4.1 and Lemma 2 of Lee, Kosorok and Fine. Lemmas 19.7 and 19.9 are Lemmas A.2 and A.1, respectively, of Lee, Kosorok and Fine, while Lemma 19.8 is Lemma 5 of Ma and Kosorok (2005a).

# 20

# Efficient Inference for Infinite-Dimensional Parameters

We now consider the special case that both $\theta$ and $\eta$ are $\sqrt{n}$ consistent in the semiparametric model $\{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$, where $\Theta \subset \mathbb{R}^k$. Often in this setting, $\eta$ may be of some interest to the data analyst. Hence, in this chapter, $\eta$ will not be considered a nuisance parameter. In the first section, we expand on the general ideas presented in the last half of Section 3.3. This includes a proof of Corollary 3.2 on Page 46. The second section presents several methods of inference, including a weighted bootstrap and a computationally efficient piggyback bootstrap, along with a brief review of a few other methods.

## 20.1 Semiparametric Maximum Likelihood Estimation

Assume that the score operator for $\eta$ has the form $h \mapsto B_{\theta,\eta}h$, where $h \in \mathcal{H}$ for some set of indices $\mathcal{H}$. The joint maximum likelihood estimator $(\hat{\theta}_n, \hat{\eta}_n)$ will then usually satisfy a Z-estimator equation $\Psi_n(\hat{\theta}_n, \hat{\eta}_n) = 0$, where $\Psi_n = (\Psi_{n1}, \Psi_{n2})$, $\Psi_{n1}(\theta, \eta) = \mathbb{P}_n \dot{\ell}_{\theta,\eta}$ and $\Psi_{n2}(\theta, \eta) = \mathbb{P}_n B_{\theta,\eta}h - P_{\theta,\eta}B_{\theta,\eta}h$, for all $h \in \mathcal{H}$. The expectation of $\Psi_n$ under the true parameter value is $\Psi = (\Psi_1, \Psi_2)$ as defined in the paragraph preceding Corollary 3.2. Here is the proof of this corollary:

**Proof of Corollary 3.2 (Page 46).** The conditions given in the first few lines of the corollary combined with the consistency of $(\hat{\theta}_n, \hat{\eta}_n)$ enable us to use Lemma 13.3 to conclude that

$$\sqrt{n}(\Psi_n - \Psi)(\hat{\theta}_n, \hat{\eta}_n) - \sqrt{n}(\Psi_n - \Psi)(\theta_0, \eta_0) = o_P(1),$$

where the convergence is uniform. Since the Donsker assumption on the score equation ensures $\sqrt{n}\Psi_n(\theta_0, \eta_0) \rightsquigarrow Z$, for some tight, mean zero Gaussian process $Z$, we have satisfied all of the conditions of Theorem 2.11, and thus $\sqrt{n}(\hat{\theta}_n - \theta_0, \hat{\eta}_n - \eta_0) \rightsquigarrow -\dot{\Psi}_0^{-1} Z$.

The remaining challenge is to establish efficiency. Recall that the differentiation used to obtain the score and information operators involves a smooth function $t \mapsto \eta_t(\theta, \eta)$ for which $\eta_0(\theta, \eta) = \eta$, $t$ is a scalar, and

$$B_{\theta,\eta} h(x) - P_{\theta,\eta} B_{\theta,\eta} h = \partial \ell_{\theta,\eta_t(\theta,\eta)}(x)/(\partial t)\big|_{t=0},$$

and where $\ell(\theta, \eta)(x)$ is the log-likelihood for a single observation (we are changing the notation slightly from Section 3.3). Note that this $\eta_t$ is not necessarily an approximately least-favorable submodel. The purpose of this $\eta_t$ is to incorporate the effect of a perturbation of $\eta$ in the direction $h \in \mathcal{H}$. The resulting one-dimensional submodel is $t \mapsto \psi_t \equiv (\theta + ta, \eta_t(\theta, \eta))$, with derivative

$$\frac{\partial}{\partial t}\psi_t\bigg|_{t=0} \equiv \dot{\psi}(a, h),$$

where $c \equiv (a, h) \in \mathbb{R}^k \times \mathcal{H}$ and $\dot{\psi} : \mathbb{R}^k \times \mathcal{H} \mapsto \mathbb{R}^k \times \operatorname{lin} H \equiv \mathcal{C}$ is a linear operator that may depend on the composite (joint) parameter $\psi \equiv (\theta, \eta)$. To be explicit about which tangent in $\mathcal{C}$ is being applied to the one-dimensional submodel, we will use the notation $\psi_{t,c}$, i.e., $\partial/(\partial t)|_{t=0}\, \psi_{t,c} = \dot{\psi}(c)$.

Define the abbreviated notation $U_\psi(c) \equiv a' \dot{\ell}_\psi + B_\psi h - P_\psi B_\psi h$, for $c = (a, h) \in \mathcal{C}$. Our construction now gives us that for any $c_1, c_2 \in \mathcal{C}$,

$$\begin{aligned} \dot{\Psi}(\dot{\psi}_0(c_2))(c_1) &= \frac{\partial}{\partial t} P_{\psi_0}\left[U_{\psi_{t,c_2}}(c_1)\right]\bigg|_{t=0,\psi=\psi_0} \\ &= -P_{\psi_0}\left[U_{\psi_0}(c_1) U_{\psi_0}(c_2)\right], \end{aligned} \tag{20.1}$$

where $\psi_0 \equiv (\theta_0, \eta_0)$. The first equality follows from the definition of $\dot{\Psi}$ as the derivative of the expected score under the true model. The second equality follows from viewing the likelihood as a two-dimensional submodel with variables $(s, t)$, where $s$ perturbs the model in the direction $c_1$ and $t$ perturbs the model in the direction $c_2$. The two-by-two information matrix in this context is the negative of the derivative of the expected score vector and is also equal to the expected outer-product of the score vector.

We know from the conclusion of the first paragraph of the proof that the influence function for $\hat{\psi}_n \equiv (\hat{\theta}_n, \hat{\eta}_n)$ is $\tilde{\psi} \equiv -\dot{\Psi}^{-1}[U_{\psi_0}(\cdot)]$. Thus, for any $c \in \mathcal{C}$,

$$\begin{aligned} P_{\psi_0}\left[\tilde{\psi}U_{\psi_0}(c)\right] &= P_{\psi_0}\left[\left(-\dot{\Psi}^{-1}\left[U_{\psi_0}(\cdot)\right]\right)U_{\psi_0}(c)\right] \\ &= -\dot{\Psi}^{-1}P_{\psi_0}\left[U_{\psi_0}(\cdot)U_{\psi_0}(c)\right] \\ &= -\dot{\Psi}^{-1}\left[-\dot{\Psi}(\dot{\psi_0}(c))(\cdot)\right] \\ &= \dot{\psi_0}(c). \end{aligned}$$

The first equality follows from the form of the influence function, the second equality follows from the fact that $\dot{\Psi}^{-1}$ is onto and linear, the third equality follows from (20.1), and the fourth equality follows from the properties of an inverse function. This means by the definition given in Section 18.1 that $\tilde{\psi}_0$ is the efficient influence function.

Since $\sqrt{n}(\hat{\psi}_n - \psi_0)$ is asymptotically tight and Gaussian with covariance that equals the covariance of the efficient influence function, we have by Theorem 18.3 that $\hat{\psi}_n$ is efficient.□

A key condition for Corollary 3.2 is that $\dot{\Psi}$ be continuously invertible and onto. This can be quite non-trivial to establish. We have demonstrated for the proportional odds model considered in Section 15.3 how this can be done by using Lemma 6.17.

For most semiparametric models where the joint parameter is regular, we can assume a little more structure than in the previous paragraphs. For many jointly regular models, we have that $\eta = A$, where $t \mapsto A(t)$ is restricted to a subset $H \in D[0,\tau]$ of functions bounded in total variation, where $\tau < \infty$. The composite parameter is thus $\psi = (\theta, A)$. We endow the parameter space with the uniform norm since this is usually the most useful in applications. Examples include many right-censored univariate regression models, including the proportional odds model of Section 15.3, certain multivariate survival models, and certain biased sampling models. We will give a few examples later on in this section.

The index set $\mathcal{H}$ we assume consists of all finite variation functions in $D[0,\tau]$, and we assign to $\mathcal{C} = \mathbb{R}^k \times \mathcal{H}$ the norm $\|c\|_{\mathcal{C}} \equiv \|a\| + \|h\|_v$, where $c = (a,h)$, $\|\cdot\|$ is the Euclidean norm, and $\|\cdot\|_v$ is the total variation norm on $[0,\tau]$. We let $\mathcal{C}_p \equiv \{c \in \mathcal{C} : \|c\|_{\mathcal{C}} \leq p\}$, where the inequality is strict when $p = \infty$. This is the same structure utilized in Section 15.3.4 for the proportional odds model aside from some minor changes in the notation. The full composite parameter $\psi = (\theta, A)$ can be viewed as an element of $\ell^\infty(\mathcal{C}_p)$ if we define

$$\psi(c) \equiv a'\theta + \int_0^\tau h(s)dA(s), \quad c \in \mathcal{C}_p, \quad \psi \in \Omega \equiv \Theta \times H.$$

As described in Section 15.3.4, $\Omega$ thus becomes a subset of $\ell^\infty(\mathcal{C}_p)$, with norm $\|\psi\|_{(p)} \equiv \sup_{c\in\mathcal{C}_p}|\psi(c)|$. Moreover, if $\|\cdot\|_\infty$ is the uniform norm on $\Omega$, then, for any $1 \leq p < \infty$, $\|\psi\|_\infty \leq \|\psi\|_{(p)} \leq 4p\|\psi\|_\infty$. Thus the uniform and $\|\cdot\|_{(p)}$ norms are equivalent.

For a direction $h \in \mathcal{H}$, we will perturb $A$ via the one-dimensional submodel $t \mapsto A_t(\cdot) = \int_0^{(\cdot)}(1 + th(s))dA(s)$. This means in the notation at the beginning of this section that $\dot{\psi}(c) = \left(a, \int_0^{(\cdot)} h(s)dA(s)\right)$. We now modify the score notation slightly. For any $c \in \mathcal{C}$, let

$$\begin{aligned} U[\psi](c) &= \frac{\partial}{\partial t}\ell\left(\theta + ta, A(\cdot) + t\int_0^{(\cdot)} h(s)dA(s)\right)\Bigg|_{t=0} \\ &= \frac{\partial}{\partial t}\ell\left(\theta + ta, A(\cdot)\right)\Big|_{t=0} + \frac{\partial}{\partial t}\ell\left(\theta, A(\cdot) + t\int_0^{(\cdot)} h(s)dA(s)\right)\Bigg|_{t=0} \\ &\equiv U_1[\psi](a) + U_2[\psi](h). \end{aligned}$$

Note that $B_\psi h = U_2[\psi](h)$, $\Psi_n(\psi)(c) = \mathbb{P}_n U[\psi](c)$, and $\Psi(\psi)(c) = P_0 U[\psi](c)$, where $P_0 = P_{\psi_0}$. In this context, $P_\psi U_2[\psi](h) = 0$ for all $h \in \mathcal{H}$ by definition of the maximum and under identifiability of the model. Thus we will not need the $P_\psi B_\psi h$ term used earlier in this section. It is important to note that the map $\psi \mapsto U[\psi](\cdot)$ actually has domain $\operatorname{lin}\Omega$ and range contained in $\ell^\infty(\mathcal{C})$.

We now consider properties of the second derivative of the log-likelihood. Let $\bar{a} \in \mathbb{R}^k$ and $\bar{h} \in \mathcal{H}$. For ease of exposition, we will use the somewhat redundant notation $c = (a, h) \equiv (c_1, c_2)$. We assume the following derivative structure exists and is valid for $j = 1, 2$ and all $c \in \mathcal{C}$:

$$\begin{aligned} &\frac{\partial}{\partial s}U_j[\theta + s\bar{a}, A + s\bar{h}](c_j)\Big|_{s=0} \\ &= \frac{\partial}{\partial s}U_j[\theta + s\bar{a}, A](c_j)\Big|_{s=0} + \frac{\partial}{\partial s}U_j[\theta, A + s\bar{h}](c_j)\Big|_{s=0}, \\ &\equiv \bar{a}'\hat{\sigma}_{1j}[\psi](c_j) + \int_0^\tau \hat{\sigma}_{2j}[\psi](c_j)(u)d\bar{h}(u), \end{aligned}$$

where $\hat{\sigma}_{1j}[\psi](c_j)$ is a random $k$-vector and $u \mapsto \hat{\sigma}_{2j}[\psi](c_j)(u)$ is a random function contained in $\mathcal{H}$. Denote $\sigma_{jk}[\psi] = P_0\hat{\sigma}_{jk}[\psi]$ and $\sigma_{jk} = \sigma_{jk}[\psi_0]$, for $j, k = 1, 2$, and where $P_0 = P_{\psi_0}$.

Let $\hat{\psi}_n = (\hat{\theta}_n, \hat{A}_n)$ be the maximizers of the log-likelihood. Then $\Psi_n(\hat{\psi}_n)(c) = 0$ for all $c \in \mathcal{C}$. Moreover, since $\operatorname{lin}\Omega$ is contained in $\mathcal{C}$, we have that the map $\bar{c} \in \operatorname{lin}\Omega \mapsto -\dot{\Psi}(\bar{c})(\cdot) \in \ell^\infty(\mathcal{C})$ has the form $-\dot{\Psi}(\bar{c})(\cdot) = \bar{c}(\sigma(\cdot))$, where $\sigma \equiv \sigma[\psi_0]$ and

$$\sigma[\psi](c) \equiv \begin{pmatrix} \sigma_{11}[\psi] & \sigma_{12}[\psi] \\ \sigma_{21}[\psi] & \sigma_{22}[\psi] \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$$

and, for any $c, \bar{c} \in \mathcal{C}$, $\bar{c}(c) = \bar{c}_1'c_1 + \int_0^\tau c_2(u)d\bar{c}_2(u)$. Provided $\sigma : \mathcal{C} \mapsto \mathcal{C}$ is continuously invertible and onto, we have that $\dot{\Psi} : \operatorname{lin}\Omega \mapsto \mathcal{C}$ is also continuously invertible and onto with inverse satisfying $-\dot{\Psi}^{-1}(c)(\cdot) = c(\sigma^{-1}(\cdot))$.

In this set-up, we will need the following conditions for some $p > 0$ in order to apply Corollary 3.2:

$$\{U[\psi](c) : \|\psi - \psi_0\| \leq \epsilon, c \in \mathcal{C}_p\} \text{ is Donsker for some } \epsilon > 0, \tag{20.2}$$

$$\sup_{c \in \mathcal{C}_p} P_0 \left| U[\psi](c) - U[\psi_0](c) \right|^2 \to 0, \text{ as } \psi \to \psi_0, \text{ and} \tag{20.3}$$

$$\sup_{c \in \mathcal{C}_p} \|\sigma[\psi](c) - \sigma[\psi_0](c)\|_{(p)} \to 0, \text{ as } \|\psi - \psi_0\|_{(p)} \to 0. \tag{20.4}$$

Note by Exercise 20.3.1 that (20.4) implies $\Psi$ is Fréchet-differentiable in $\ell^\infty(\mathcal{C}_p)$. It is also not hard to verify that if Conditions (20.2)–(20.4) hold for some $p > 0$, then they hold for all $0 < p < \infty$ (see Exercise 20.3.2). This yields the following corollary, whose simple proof is saved as an exercise (see Exercise 20.3.3):

COROLLARY 20.1 *Assume Conditions (20.2)–(20.4) hold for some $p > 0$, that $\sigma : \mathcal{C} \mapsto \mathcal{C}$ is continuously invertible and onto, and that $\hat{\psi}_n$ is uniformly consistent for $\psi_0$ with*

$$\sup_{c \in \mathcal{C}_1} \left| \mathbb{P}_n \Psi_n(\hat{\psi}_n)(c) \right| = o_{P_0}(n^{-1/2}).$$

*Then $\hat{\psi}_n$ is efficient with*

$$\sqrt{n}(\hat{\psi}_n - \psi_0)(\cdot) \rightsquigarrow Z(\sigma^{-1}(\cdot))$$

*in $\ell^\infty(\mathcal{C}_1)$, where $Z$ is the tight limiting distribution of $\sqrt{n}\mathbb{P}_n U[\psi_0](\cdot)$.*

Note that we actually need $Z$ to be a tight element in $\ell^\infty(\sigma^{-1}(\mathcal{C}_1))$, but the linearity of $U[\psi](\cdot)$ ensures that if $\sqrt{n}\mathbb{P}_n U[\psi_0](\cdot)$ converges to $Z$ in $\ell^\infty(\mathcal{C}_1)$, then it will also converge weakly in $\ell^\infty(\mathcal{C}_p)$ for any $p < \infty$. Since $\sigma$ is continuously invertible by assumption, there exists a $p_0 < \infty$ such that

$$\sup_{c \in \mathcal{C}_1} \left| \sigma^{-1}(c) \right| \leq p_0,$$

and thus $\sigma^{-1}(\mathcal{C}_1) \subset \mathcal{C}_{p_0}$.

This kind of analysis was used extensively in the proportional odds model example studied in Section 15.3. In fact, the maximum likelihood joint estimator $(\hat{\theta}_n, \hat{A}_n)$ studied in Section 15.3 can be easily shown to satisfy the conditions of Corollary 20.1 (see Exercise 20.3.4). Thus $(\hat{\theta}_n, \hat{A}_n)$ is asymptotically efficient for estimating $(\theta_0, A_0)$ in the proportional odds example. Since the $\|\cdot\|_{(1)}$ norm dominates the uniform norm, the weak convergence and efficiency results we obtain with respect to the $\|\cdot\|_{(1)}$ norm will carry through for the uniform norm.

The next section, Section 20.2, will study methods of inference for maximum likelihood settings that satisfy the conditions of Corollary 20.1. Typically, these maximum likelihood settings involve an empirical likelihood

structure for the infinite-dimensional component, as discussed at the beginning of Section 3.3. We will consider three examples in detail: the Cox regression model for right-censored data, the proportional odds model for right-censored data, and an interesting biased sampling model. Two additional models were also studied using the methods of this chapter in Dixon, Kosorok and Lee (2005), but we will not consider them further here: the odds-rate regression model for right-censored survival data and the correlated gamma frailty regression model for multivariate right-censored survival data. Before moving on to Section 20.2, we briefly discuss the application of Corollary 20.1 to the Cox model and the biased sampling examples. The application of Corollary 20.1 to the proportional odds example has already been discussed.

*The Cox model for right-censored data.*

This model has been explored extensively in previous sections, and both weak convergence and efficiency have already been established. Nevertheless, it is useful to study this model again from the perspective of this section. Accordingly, we will let $\theta$ be the regression effect and $A$ the baseline hazard, with the observed data $X = (U, \delta, Z)$, where $U$ is the minimum of the failure time $T$ and censoring time $C$, and where $\delta$ and $Z$ have the unusual definitions. We make the usual assumptions for this model as done in Section 4.2.2, including requiring the baseline hazard to be continuous, except that we will use $(\theta, A)$ to denote the model parameters $(\beta, \Lambda)$.

It is not hard to verify that $U_1[\psi](a) = \int_0^\tau Z' a dM_\psi(s)$ and $U_2[\psi](h) = \int_0^\tau h(s) dM_\psi(s)$, where $M_\psi(t) \equiv N(t) - \int_0^t Y(s) e^{\theta' Z} dA(s)$ and $N$ and $Y$ are the usual counting and at-risk processes. It is also easy to show that the components of $\sigma$ are defined by

$$\begin{aligned}
\sigma_{11} a &= \int_0^\tau P_0 \left[ Z Z' Y(s) e^{\theta_0' Z} \right] dA_0(s) a, \\
\sigma_{12} h &= \int_0^\tau P_0 \left[ Z Y(s) e^{\theta_0' Z} \right] h(s) dA_0(s), \\
\sigma_{21} a &= P_0 \left[ Z' Y(\cdot) e^{\theta_0' Z} \right] a, \text{ and} \\
\sigma_{22} h &= P_0 \left[ Y(\cdot) e^{\theta_0' Z} \right] h(\cdot).
\end{aligned}$$

The maximum likelihood estimator is $\hat{\psi}_n = (\hat{\theta}_n, \hat{A}_n)$, where $\hat{\theta}_n$ is the maximizer of the well-known partial likelihood and $\hat{A}_n$ is the Breslow estimator. We will now establish that the conditions of Corollary 20.1 hold for this example. Conditions (20.2)–(20.4) are easy to verify by recycling arguments we have used previously. Likewise, most of the remaining conditions are easy to verify.

The only somewhat difficult condition to verify is that $\sigma$ is continuously invertible and onto. We will use Lemma 6.17 to do this. First, it is

easy to verify that $\sigma = \kappa_1 + \kappa_2$, where $\kappa_1 c \equiv (a, \rho_0(\cdot)h(\cdot))$, $\kappa_2 \equiv \sigma - \kappa_1$, $\rho_0(t) \equiv P_0[Y(t)e^{\theta_0' Z}]$, and where $\kappa_2$ is a compact operator and $\kappa_1$ is trivially continuously invertible and onto. Lemma 6.17 will now imply that $\sigma$ is continuously invertible and onto, provided we can show that $\sigma$ is one-to-one.

Accordingly, let $c \in \mathcal{C}$ be such that $\sigma(c) = 0$, where, as usual, $c = (a, h)$. Then we also have that $\bar{c}(\sigma(c)) = 0$, where $\bar{c} = (a, \int_0^t h(s)dA_0(s))$. This implies, after a little algebra, that

$$\begin{aligned} 0 &= \int_0^\tau P_0 \left( [a'Z + h(s)]^2 Y(s) e^{\theta_0' Z} \right) dA_0(s) \\ &= \int_0^\tau P_0 \left( [a'Z + h(s)]^2 dN(s) \right), \end{aligned}$$

which implies that

$$[a'Z + h(T)]^2 \delta = 0, \quad P_0\text{-almost surely.} \tag{20.5}$$

This now forces $a = 0$ and $h = 0$ (see Exercise 20.3.5). Thus $\sigma$ is one-to-one. Hence all of the conditions of Corollary 20.1 hold, and we conclude that $\hat{\psi}_n$ is efficient.

*A biased sampling model.*

We will now consider a special case of a class of biased sampling models which were studied by Gilbert (2000). The data consists of $n$ i.i.d. realizations of $X = (\delta, Y)$. Here, $\delta \in \{0, 1\}$ is a random stratum identifier, taking on the value $j$ with selection probability $\lambda_j > 0$, $j = 0, 1$, with $\lambda_0 + \lambda_1 = 1$. Given $\delta = j$, $Y \in [0, \tau]$ has distribution $F_j$ defined on a sigma field of subsets $\mathcal{B}$ of $[0, \tau]$ by $F_j(B, \theta, A) \equiv W_j^{-1}(\theta, A) \int_B w_j(u, \theta) dA(u)$ for $B \in \mathcal{B}$. The $w_j$, $j = 0, 1$, are nonnegative (measurable) stratum weight functions assumed to be known up to the finite dimensional parameter $\theta \in \Theta \subset \mathbb{R}$. We will assume hereafter that $w_0(t, \theta) = 1$ and that $w_1(t, \theta) = e^{\theta t}$. $W_j(\theta, A) \equiv \int_0^\tau w_j(u, \theta) dA(u)$ is assumed to be finite for all $\theta \in \Theta$. The probability measure $A$ is the unknown infinite dimensional parameter of interest, and $\psi = (\theta, A)$ is the joint parameter. We assume that $A_0$ is continuous with support on all of $[0, \tau]$. The goal is to estimate $\psi$ based on information from samples from the $F_j$ distributions, $j = 0, 1$. Thus the log-likelihood for a single observation is

$$\begin{aligned} \ell(\psi)(X) &= \log w_\delta(Y, \theta) + \log \Delta A(Y) - \log W_\delta(\theta, A), \\ &= \delta \theta Y + \log \Delta A(Y) - \log \int_0^\tau e^{\delta \theta s} dA(s), \end{aligned}$$

where $\Delta A(Y)$ is the probability mass of $A$ at $Y$.

Thus the score functions are

$$
\begin{aligned}
U_1[\psi](a) &= \delta\left[Y - \frac{\int_0^\tau s e^{\delta\theta s} dA(s)}{\int_0^\tau e^{\delta\theta s} dA(s)}\right] a, \quad \text{and} \\
U_2[\psi](h) &= h(Y) - \frac{\int_0^\tau e^{\delta\theta s} h(s) dA(s)}{\int_0^\tau e^{\delta\theta s} dA(s)}.
\end{aligned}
$$

The components of $\sigma$ are obtained by taking the expectations under the true distribution $P_0$ of $\hat{\sigma}_{jk}$, $j,k = 1,2$, where

$$
\begin{aligned}
\hat{\sigma}_{11} a &= \left(E_\delta([\delta y]^2) - [E_\delta(\delta y)]^2\right) a, \\
\hat{\sigma}_{21} a &= \frac{e^{\delta\theta_0(\cdot)}}{\int_0^\tau e^{\delta\theta_0 s} dA_0(s)} \left[\delta(\cdot) - E_\delta(\delta y)\right] a, \\
\hat{\sigma}_{12} h &= E_\delta(\delta y h(y)) - E_\delta(\delta y) E_\delta(h(y)), \quad \text{and} \\
\hat{\sigma}_{22} h &= \frac{e^{\delta\theta_0(\cdot)}}{\int_0^\tau e^{\delta\theta_0 s} dA_0(s)} \left[h(\cdot) - E_\delta(h(y))\right],
\end{aligned}
$$

where, for a $\mathcal{B}$-measurable function $y \mapsto f(y)$,

$$
E_j(f(y)) \equiv \frac{\int_0^\tau f(y) e^{j\theta_0 y} dA_0(y)}{\int_0^\tau e^{j\theta_0 y} dA_0(y)},
$$

for $j = 0, 1$. Then $\sigma_{jk} = P_0 \hat{\sigma}_{jk}$, $j,k = 1,2$.

We now show that $\sigma : \mathcal{C} \mapsto \mathcal{C}$ is continuously invertible and onto. First, as with the previous example, it is easy to verify that $\sigma = \kappa_1 + \kappa_2$, where $\kappa_1 c \equiv (a, \rho_0(\cdot) h(\cdot))$, $\kappa_2 \equiv \sigma - \kappa_1$,

$$
\rho_0(\cdot) \equiv P_0\left[\frac{e^{\delta\theta_0(\cdot)}}{\int_0^\tau e^{\delta\theta_0 s} dA_0(s)}\right],
$$

and where $\kappa_2$ is a compact operator and $\kappa_1$ is continuously invertible and onto. Provided we can show that $\sigma$ is one-to-one, we will be able to utilize again Lemma 6.17 to obtain that $\sigma$ is continuously invertible and onto.

Our argument for showing that $\sigma$ is one-to-one will be similar to that used for the Cox model for right censored data. Accordingly, let $c = (a, h) \in \mathcal{C}$ satisfy $\sigma c = 0$, and let $\bar{c} = (a, \int_0^{(\cdot)} h(s) dA_0(s))$. Thus $\bar{c}(\sigma c) = 0$. After some algebra, it can be shown that this implies

$$
\begin{aligned}
0 &= P_0 E_\delta \left(a\delta y + h(y) - E_\delta[a\delta y + h(y)]\right)^2 \\
&= \lambda_0 V_0(h(y)) + \lambda_1 V_1(ay + h(y)),
\end{aligned} \tag{20.6}
$$

where, for a measurable function $y \mapsto f(y)$, $V_j(f(y))$ is the variance of $f(Y)$ given $\delta = j$, $j = 0, 1$. Recall that both $\lambda_0$ and $\lambda_1$ are positive. Thus, since (20.6) implies $V_0(h(y)) = 0$, $y \mapsto h(y)$ must be a constant function. Since (20.6) also implies $V_1(ay + h(y)) = 0$, we now have that $a = 0$. Hence

$h(Y) = E_\delta(h(y))$ almost surely. Thus $P_0 h^2(Y) = 0$, which implies $h = 0$ almost surely. Hence $c = 0$, and thus $\sigma$ is one-to-one.

Conditions (20.2)–(20.4) can be established for $p = 1$ by recycling previous arguments (see Exercise 20.3.6). Gilbert (2000) showed that the maximum likelihood estimator $\hat{\psi}_n = \text{argmax}_\psi \mathbb{P}_n l_\psi(X)$ is uniformly consistent for $\psi_0$. Since $\Psi_n(\hat{\psi}_n)(c) = 0$ almost surely for all $c \in \mathcal{C}$ by definition of the maximum, all of the conditions of Corollary 20.1 hold for $\hat{\psi}_n$. Thus $\hat{\psi}_n$ is efficient.

## 20.2 Inference

We now present several methods of inference for the semiparametric maximum likelihood estimation setting of Corollary 20.1. The first method is the specialization of the bootstrap approach for Z-estimators described in Section 13.2.3 to the current setting. The second method, the "piggyback bootstrap," is a much more computationally efficient method that takes advantage of the special structure of efficient estimators. We conclude this section with a brief review of several additional methods of inference for regular, infinite-dimensional parameters.

### *20.2.1 Weighted and Nonparametric Bootstraps*

Recall the nonparametric and weighted bootstrap methods for Z-estimators described in Section 13.2.3. Let $\mathbb{P}_n^\circ$ and $\mathbb{G}_n^\circ$ be the bootstrapped empirical measure and process based on either kind of bootstrap, and let $\overset{\text{P}}{\underset{\circ}{\rightsquigarrow}}$ denote either $\overset{\text{P}}{\underset{W}{\rightsquigarrow}}$ for the nonparametric version or $\overset{\text{P}}{\underset{\xi}{\rightsquigarrow}}$ for the weighted version. We will use $\hat{\psi}_n^\circ$ to denote an approximate maximizer of the bootstrapped empirical log-likelihood $\psi \mapsto \mathbb{P}_n^\circ \ell(\psi)(X)$, and we will denote $\Psi_n^\circ(\psi)(c) \equiv \mathbb{P}_n^\circ U[\psi](c)$ for all $\psi \in \Omega$ and $c \in \mathcal{C}$. We now have the following simple corollary, where $\mathcal{X}_n$ is the $\sigma$-field of the observations $X_1, \ldots, X_n$:

COROLLARY 20.2 *Assume the conditions of Corollary 20.1, and, in addition, that* $\hat{\psi}_n^\circ \overset{\text{as*}}{\to} \psi_0$ *unconditionally and*

$$\text{P}\left( \sqrt{n} \sup_{c \in \mathcal{C}_1} |\Psi_n(\hat{\psi}_n^\circ)(c)| \,\middle|\, \mathcal{X}_n \right) \quad = \quad o_P(1). \tag{20.7}$$

*Then the conclusions of Corollary 20.1 hold and* $\sqrt{n}(\hat{\psi}_n^\circ - \hat{\psi}_n) \overset{\text{P}}{\underset{\circ}{\rightsquigarrow}} Z(\sigma^{-1}(\cdot))$ *in* $\ell^\infty(\mathcal{C}_1)$*, i.e., the limiting distribution of* $\sqrt{n}(\hat{\psi}_n - \psi_0)$ *and the conditional limiting distribution of* $\sqrt{n}(\hat{\psi}_n^\circ - \hat{\psi}_n)$ *given* $\mathcal{X}_n$ *are the same.*

Before giving the simple proof of this corollary, we make a few remarks. First, we make the obvious point that if $\hat{\psi}_n^\circ$ is the maximizer of $\psi \mapsto$

$\mathbb{P}_n^\circ \ell(\psi)(X)$, then (20.7) is easily satisfied by the definition of the maximizer. Second, we note that method of proving consistency of $\hat{\psi}_n^\circ$ is different than that used in Theorem 13.4, which utilizes the identifiability properties of the limiting estimating equation $\Psi$ as expressed in Condition (13.1). The reason for this is that for many semiparametric models, identifiability is quite tricky to establish and for many complicated models it is not clear how to prove identifiability of $\Psi$. Establishing consistency of $\hat{\psi}_n^\circ$ directly is often more fruitful.

For the weighted bootstrap, maximizing $\psi \mapsto \mathbb{P}_n^\circ \ell(\psi)(X)$ is equivalent to maximizing $\psi \mapsto \mathbb{P}_n \xi \ell(\psi)(X)$, where the positive weights $\xi_1, \ldots, \xi_n$ are i.i.d. and independent of the data $X_1, \ldots, X_n$. By Corollary 9.27, $\{\xi \ell(\psi)(X) : \psi \in \Omega\}$ forms a Glivenko-Cantelli class with integrable envelope if and only if $\{\ell(\psi)(X) : \psi \in \Omega\}$ is Glivenko-Cantelli with integrable envelope. Thus many of the consistency arguments used in establishing consistency for $\hat{\psi}_n$ can be applied—almost without modification—to verifying unconditional consistency of $\hat{\psi}_n^\circ$. This is true, for example, with the existence and consistency arguments for the joint estimator from the proportional odds model for right-censored data discussed in Sections 15.3.2 and 15.3.3. Thus, for this proportional odds example, $\hat{\psi}_n^\circ$ is unconditionally consistent and satisfies (20.7), and thus the weighted bootstrap is valid in this case. It is not clear what arguments would be needed to obtain the corresponding results for the nonparametric bootstrap.

An important advantage of the weighted and nonparametric bootstraps is that they do not require the likelihood model to be correctly specified. An example of the validity of these methods under misspecification of univariate frailty regression models for right-censored data, a large class of models which includes both the Cox model and the proportional odds model as a special cases, is described in Kosorok, Lee and Fine (2004). An important disadvantage of these bootstraps is that they are very computationally intense since the weighted likelihood must be maximized over the entire parameter space for every bootstrap realization. This issue will be addressed in Section 20.2.2 below.

**Proof of Corollary 20.2.** The proof essentially follows from the fact that the conditional bootstrapped distribution of a Donsker class is automatically consistent via Theorem 2.6. Since conditional weak convergence implies unconditional weak convergence (as argued in the proof of Theorem 10.4), both Lemma 13.3 and Theorem 2.11 apply to $\Psi_n^\circ$, and thus

$$\sup_{c \in \mathcal{C}_p} \left| \sqrt{n}(\hat{\psi}_n^\circ - \psi_0)(\sigma(c)) - \sqrt{n}(\Psi_n^\circ - \Psi)(c) \right| = o_{P_0}(1),$$

unconditionally, for any $0 < p < \infty$. Combining this with previous results for $\hat{\psi}_n$, we obtain for any $0 < p < \infty$

$$\sup_{c \in \mathcal{C}_p} \left| \sqrt{n}(\hat{\psi}_n^\circ - \hat{\psi}_n)(\sigma(c)) - \sqrt{n}(\Psi_n^\circ - \Psi_n)(c) \right| = o_{P_0}(1).$$

Since $\{U[\psi_0](c) : c \in \mathcal{C}_p\}$ is Donsker for any $0 < p < \infty$, we have the desired conclusion by reapplication of Theorem 2.6 and the continuous invertibility of $\sigma$.□

### *20.2.2 The Piggyback Bootstrap*

We now discuss an inferential method for the joint maximum likelihood estimator $\hat{\psi}_n = (\hat{\theta}_n, \hat{A}_n)$ that is computationally much more efficient than the usual bootstrap. From the previous chapter, Chapter 19, we have learned a method—the "profile sampler"—for generating random realizations $\theta_n$ such that $\sqrt{n}(\theta_n - \hat{\theta}_n)$ given the data has the same limiting distribution as $\sqrt{n}(\hat{\theta}_n - \theta_0)$ does unconditionally. The *piggyback bootstrap* will utilize these $\theta_n$ realizations to improve computational efficiency. It turns out that it does not matter how the $\theta_n$ are generated, provided $\sqrt{n}(\theta_n - \hat{\theta}_n)$ has the desired conditional limiting distribution. For any $\theta \in \Theta$, let $\hat{A}^\circ_\theta = \text{argmax}_A \mathbb{P}^\circ_n \ell(\theta, A)(X)$, where $\mathbb{P}^\circ_n$ is the weighted bootstrap empirical measure. In this section, we will not utilize the nonparametric bootstrap but will only consider the weighted bootstrap.

The main idea of the piggyback bootstrap of Dixon, Kosorok and Lee (2005) is to generate a realization of $\theta_n$, then generate the random weights $\xi_1, \ldots, \xi_n$ in $\mathbb{P}^\circ_n$ independent of both the data and $\theta_n$, and then compute $\hat{A}^\circ_{\theta_n}$. This generates a joint realization $\hat{\psi}^\circ_n \equiv (\theta_n, \hat{A}^\circ_n)$. For instance, one can generate a sequence of $\theta_n$s, $\theta_n^{(1)}, \ldots, \theta_n^{(m)}$, using the profile sampler. For each $\theta_n^{(j)}$, a new draw of the random weights $\xi_1, \ldots, \xi_n$ is made, and $\hat{A}^\circ_{(j)} \equiv \hat{A}^\circ_{\theta_n^{(j)}}$ is computed. A joint realization $\hat{\psi}^\circ_{(j)} \equiv (\theta_n^{(j)}, \hat{A}^\circ_{\theta_n^{(j)}})$ is thus obtained for each $j = 1, \ldots, m$.

Under regularity conditions which we will delineate, the conditional distribution of $\sqrt{n}(\hat{\psi}^\circ_n - \hat{\psi}_n)$ converges to the same limiting distribution as $\sqrt{n}(\hat{\psi}_n - \psi_0)$ does unconditionally. Hence the realizations $\hat{\psi}^\circ_{(1)}, \ldots, \hat{\psi}^\circ_{(m)}$ can be used to construct joint confidence bands for Hadamard-differentiable functions of $\psi_0 = (\theta_0, A_0)$, as permitted by Theorem 12.1. For example, this could be used to construct confidence bands for estimated survival curves from a proportional odds model for a given covariate value.

We now present a few regularity conditions which, in addition to the assumptions of Corollary 20.1, are sufficient for the piggyback bootstrap to be valid. First, decompose $Z$ from Corollary 20.1 into two parts, $(Z_1, Z_2) = Z$, where $Z_1 \in \mathbb{R}^k$ and $Z_2 \in \ell^\infty(\mathcal{H}_p)$, for some $0 < p < \infty$. Let $M$ denote the random quantities used to generate $\theta_n$, so that $\sqrt{n}(\theta_n - \hat{\theta}_n) \underset{M}{\overset{\text{P}}{\rightsquigarrow}} Z_1'$, where $Z_1'$ is a Gaussian vector independent of $Z$ with mean zero and covariance equal to the upper $k \times k$ entry of $\sigma^{-1}$; and let $\xi$ denote $\xi_1, \xi_2, \ldots$, so that $\underset{\xi}{\overset{\text{P}}{\rightsquigarrow}}$ will signify conditional convergence of the weighted bootstrap process in probability, given the data. We assume the random quantities represented

by $M$ are independent of those represented by $\xi$, i.e., the generation of $\theta_n$ is independent of the random weights $\xi_1, \xi_2, \ldots$. For simplicity of exposition, we also require $\hat{\psi}_n = (\hat{\theta}_n, \hat{A}_n)$ to be a maximizer of $\psi \mapsto \mathbb{P}_n \ell(\psi)(X)$. We also need that $\sigma_{22} : \mathcal{H}_p \mapsto \mathcal{H}_p$ is continuously invertible and onto; and, for any potentially random sequence $\tilde{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$, $\hat{A}^{\circ}_{\tilde{\theta}_n} \xrightarrow{\mathrm{P}} A_0$ in $\ell^{\infty}(\mathcal{H}_p)$, unconditionally, for some $0 < p < \infty$. We now present the main result of this section:

COROLLARY 20.3 *Assume the above conditions in addition to the conditions of Corollary 20.1. Then the conclusions of Corollary 20.1 hold and*

$$\sqrt{n} \begin{pmatrix} \theta_n - \hat{\theta}_n \\ \hat{A}^{\circ}_{\theta_n} - \hat{A}_n \end{pmatrix} \overset{\mathrm{P}}{\underset{M,\xi}{\rightsquigarrow}} Z(\sigma^{-1}(\cdot)), \text{ in } \ell^{\infty}(\mathcal{C}_1).$$

Before giving the proof, we note several things. First, for the two examples considered in detail at the end of Section 20.1, continuous invertibility of $\sigma_{22}$ is essentially implied by the assumed structure and continuous invertibility of $\sigma$. To see this, it is enough to verify that $\sigma_{22}$ is one-to-one. Accordingly, fix an $h \in \mathcal{H}$ that satisfied $\sigma_{22} h = 0$. We thus have that $\bar{c}(\sigma c) = 0$, where $\bar{c} = \left(0, \int_0^{(\cdot)} h(u) dA_0(u)\right)$ and $c = (0, h)$. For both of the examples, we verified that this implies $c = 0$, which obviously implies $h = 0$. While the above argument does not directly work for the proportional odds example, it is not hard to verify the desired result by using simple, recycled arguments from the proof of Theorem 15.9 (see Exercise 20.3.7). In general, establishing continuous invertibility of $\sigma_{22}$ is quite easy to do, and is almost automatic, once continuous invertibility of $\sigma$ has been established.

The only challenging condition to establish is $\hat{A}^{\circ}_{\tilde{\theta}_n} \xrightarrow{\mathrm{P}} A_0$ for any sequence $\tilde{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$. We will examine this, after the proof given below of the above corollary, for three examples, both the Cox and proportional odds models for right censored data and the biased sampling model considered at the end of Section 20.1.

As a final note, we point out that the computational advantages of the piggy-back bootstrap over the usual bootstrap can be quite dramatic. A simulation study given in Section 4 of Dixon, Kosorok and Lee (2005) identified a 30-fold decrease in computation time (as measure by both actual time and number of maximizations performed) from using the piggyback bootstrap instead of the weighted bootstrap.

**Proof of Corollary 20.3.** Similar arguments to those used in Exercise 20.3.2 can be used to verify that the assumptions of the corollary which apply for some $0 < p < \infty$ will hold for all $0 < p < \infty$. We will therefore assume without loss of generality that $p$ is large enough for both $\sigma^{-1}(\mathcal{C}_1) \in \mathcal{C}_p$ and $\sigma_{22}^{-1}(\mathcal{H}_1) \in \mathcal{H}_p$. Recycling arguments used before, we can readily obtain that

$$\sqrt{n}(\mathbb{P}_n^{\circ} - \mathbb{P}_n) U_2[\theta_n, \hat{A}^{\circ}_{\theta_n}](\cdot) = \sqrt{n}(\mathbb{P}_n^{\circ} - \mathbb{P}_n) U_2[\theta_0, A_0](\cdot) + o_{P_0}^{\mathcal{H}_p}(1),$$

unconditionally. However, the left-hand-side of this expression equals, by the definition of a maximizer,

$$\begin{aligned} -\sqrt{n}\mathbb{P}_n U_2[\theta_n, \hat{A}^\circ_{\theta_n}] &= -\sqrt{n}\mathbb{P}_n \left( U_2[\theta_n, \hat{A}^\circ_{\theta_n}] - U_2[\hat{\theta}_n, \hat{A}_n] \right) \\ &= -\sqrt{n}P \left( U_2[\theta_n, \hat{A}^\circ_{\theta_n}] - U_2[\hat{\theta}_n, \hat{A}_n] \right) + o^{\mathcal{H}_p}_{P_0}(1), \\ &= -\sigma_{21}\sqrt{n}(\theta_n - \hat{\theta}_n) - \sigma_{22}\sqrt{n}(\hat{A}^\circ_{\theta_n} - \hat{A}_n) + o^{\mathcal{H}_p}_{P_0}(1), \end{aligned}$$

where the second equality follows the stochastic equicontinuity assured by Condition (20.3), and the third equality follows from the Fréchet differentiability of $\Psi$ as assured by Exercise 20.3.1.

Thus we have, unconditionally, that $\sqrt{n}(\hat{A}^\circ_{\theta_n} - \hat{A}_n) = -\sigma_{22}^{-1}\sigma_{21}\sqrt{n}(\theta_n - \hat{\theta}_n) - \sqrt{n}(\mathbb{P}^\circ_n - \mathbb{P}_n)U_2[\psi_0](\sigma_{22}^{-1}(\cdot)) + o^{\mathcal{H}_p}_{P_0}(1)$. Hence

$$\sqrt{n}\begin{pmatrix} \theta_n - \hat{\theta}_n \\ \hat{A}^\circ_{\theta_n} - \hat{A}_n \end{pmatrix} \overset{\mathrm{P}}{\underset{M,\,\xi}{\rightsquigarrow}} \begin{pmatrix} Z_1' \\ Z_2(\sigma_{22}^{-1}(\cdot)) - \sigma_{22}^{-1}\sigma_{21}Z_1' \end{pmatrix},$$

where $Z_1'$ and $Z_2$ are independent mean-zero Gaussian processes with covariances defined above. Hence the right-hand-side has joint covariance $\bar{c}'V_0c$, where $\bar{c}' \equiv \left(\bar{c}_1, \int_0^{(\cdot)} \bar{c}_2(s)dA_0(s)\right)$, and $\bar{c} \equiv (\bar{c}_1, \bar{c}_2)$, for any $\bar{c}, c \in \mathcal{C}_1$, and where

$$V_0 \equiv \begin{pmatrix} v_{11} & -v_{11}\sigma_{12}\sigma_{22}^{-1} \\ -\sigma_{22}^{-1}\sigma_{21}v_{11} & \sigma_{22}^{-1} + \sigma_{22}^{-1}\sigma_{21}v_{11}\sigma_{12}\sigma_{22}^{-1} \end{pmatrix},$$

and

$$v_{11} \equiv \left(\sigma_{11} - \sigma_{12}\sigma_{22}^{-1}\sigma_{21}\right)^{-1}.$$

Since $V_0 = \sigma^{-1}$, the desired result follows.□

*The Cox model under right censoring.*

All that remains to show for the Cox model is that $\hat{A}^\circ_{\tilde{\theta}_n}$ converges uniformly to $A_0$, unconditionally, for any sequence $\tilde{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$. Fortunately, the maximizer of $A \mapsto \mathbb{P}^\circ_n \ell(\tilde{\theta}_n, A)(X)$ has the explicit "Breslow estimator" form:

$$\hat{A}^\circ_{\tilde{\theta}_n}(t) = \int_0^t \frac{\mathbb{P}^\circ_n dN(s)}{\mathbb{P}^\circ_n Y(s)e^{\tilde{\theta}_n' Z}}.$$

Since the class of functions $\{Y(t)e^{\theta' Z} : t \in [0,\tau], \|\theta - \theta_0\| \le \epsilon\}$ is Glivenko-Cantelli for some $\epsilon > 0$, $\mathbb{P}^\circ_n Y(t)e^{\tilde{\theta}_n' Z} \xrightarrow{\mathrm{P}} P_0 Y(t)e^{\theta_0' Z}$ uniformly and unconditionally. Since the class $\{N(t) : t \in [0,\tau]\}$ is also Glivenko-Cantelli, and since the map $(U,V) \mapsto \int (1/V)dU$ is continuous—provided $V$ is bounded below—via Lemma 12.3, we obtain the desired result that $\hat{A}^\circ_{\tilde{\theta}_n} \xrightarrow{\mathrm{P}} A_0$ uniformly and unconditionally. Thus the conclusions of Corollary 20.3 hold.

The piggyback bootstrap in this instance is incredibly simple to implement. Let

$$\hat{V} \equiv \int_0^\tau \left[ \frac{\mathbb{P}_n Z Z' Y(t) e^{\hat{\theta}_n' Z}}{\mathbb{P}_n Y(t) e^{\hat{\theta}_n' Z}} - \left\{ \frac{\mathbb{P}_n Z Y(t) e^{\hat{\theta}_n' Z}}{\mathbb{P}_n Y(t) e^{\hat{\theta}_n' Z}} \right\}^{\otimes 2} \right] \mathbb{P}_n dN(t),$$

and let $\theta_n = \hat{\theta}_n + Z_n/\sqrt{n}$ where $Z_n$ is independent of $\xi_1, \ldots, \xi_n$ and normally distributed with mean zero and variance $\hat{V}$. Then $\hat{\psi}_n^\circ = (\theta_n, \hat{A}_{\theta_n}^\circ)$ will be an asymptotically valid Monte Carlo replication of the correct joint limiting distribution of $\hat{\psi}_n = (\hat{\theta}_n, \hat{A}_n)$.

*The proportional odds model under right censoring.*

A careful inspection of the arguments for estimator existence given in Section 15.3.2 yields the conclusion that $\limsup_{n\to\infty} \hat{A}_{\theta_n'}^\circ(\tau) < \infty$ with inner probability one, unconditionally, for any arbitrary (not necessarily convergent) deterministic sequence $\theta_n' \in \Theta$. This follows primarily from the fact that for any Glivenko-Cantelli class $\mathcal{F}$, $\|\mathbb{P}_n^\circ - \mathbb{P}_n\|_\mathcal{F} \overset{\text{as*}}{\to} 0$ via Corollary 10.14.

Similar reasoning can be applied to the arguments of Section 15.3.3 to establish that $\hat{A}_{\theta_n'}^\circ \overset{\text{as*}}{\to} A_0$, unconditionally, for any deterministic sequence $\theta_n' \to \theta_0$. Thus, for any arbitrary deterministic sequence of sets $\Theta_n' \ni \theta_0$ converging to $\{\theta_0\}$, we have $\sup_{\theta \in \Theta_n'} \|\hat{A}_n^\circ(\theta) - A_0\|_\infty \overset{\text{as*}}{\to} 0$. This yields that $\|\hat{A}_{\tilde{\theta}_n}^\circ - A_0\|_\infty \overset{\text{P}}{\to} 0$ for any sequence $\tilde{\theta}_n \overset{\text{P}}{\to} \theta_0$. Hence the conclusions of Corollary 20.3 follow, provided we can obtain random draws $\theta_n$ which satisfy the requisite conditions.

We will show in Chapter 22 that the profile sampler is valid for the proportional odds model under right censoring, and thus $\theta_n$ values generated from the profile sampler will have the needed properties. Hence we will be able to generate random realizations $(\theta_n, \hat{A}_{\theta_n}^\circ)$ which have the desired conditional joint limiting distribution and thus can be used for inference. In this instance, one can adapt the arguments of Section 15.3.1 to obtain a computationally efficient stationary point algorithm which involves solving iteratively

$$\hat{A}_{\theta_n}^\circ(t) = \int_0^t \left\{ \mathbb{P}_n^\circ W(s; \hat{\psi}_{\theta_n}^\circ) \right\}^{-1} \mathbb{P}_n^\circ dN(s),$$

where

$$W(t; \psi) \equiv \frac{(1+\delta) e^{\theta' Z} Y(t)}{1 + e^{\theta' Z} A(U)},$$

$\psi = (\theta, A)$ and $\hat{\psi}_{\theta_n}^\circ = (\theta_n, \hat{A}_{\theta_n}^\circ)$. This will result in significant computational savings over the usual bootstrap.

*A biased sampling model.*

Establishing the regularity conditions for the biased sampling model presented at the end of Section 20.1 is somewhat involved, and we will omit the

proofs which are given in Dixon, Kosorok and Lee (2005). However, we will discuss implementation of the piggyback bootstrap for this model. Firstly, we can use arguments given in Vardi (1985) to obtain that the maximum likelihood estimator $\hat{\theta}_n$ is the maximizer of the partial log-likelihood

$$p\ell_n(\theta) \equiv \sum_{j=0,1} \lambda_{nj} \int_0^\tau \log \left[ \frac{w_j(u,\theta) B_n^{-j}(\theta)}{\sum_{k=0,1} \lambda_{nk} w_k(u,\theta) B_n^{-k}(\theta)} \right] dG_{jk}(u),$$

where $\lambda_{nj} \equiv n_j/n$, $n_j$ is the number of observations in stratum $j$, $G_{nj}$ is the empirical distribution of the observations $Y$ in stratum $j$, for $j = 0, 1$, $B_n(\theta)$ is the unique solution of

$$\int_0^\tau \frac{w_1(u,\theta)}{B_n(\theta)\lambda_{n0} w_0(u,\theta) + \lambda_{n1} w_1(u,\theta)} dG_n(u) = 1,$$

and where $G_n \equiv \lambda_{n0} G_{n0} + \lambda_{n1} G_{n1}$. Thus we can obtain draws $\theta_n$ via the profile sampler applied to $p\ell_n(\theta)$.

Secondly, arguments in Vardi (1985) can also be used to obtain a closed form solution for $\hat{A}_\theta$:

$$\hat{A}_\theta(t) = \frac{\int_0^t \left[ \sum_{j=0,1} \lambda_{nj} w_j(u,\theta) B_n^{-j}(\theta) \right]^{-1} dG_n(u)}{\int_0^\tau \left[ \sum_{j=0,1} \lambda_{nj} w_j(s,\theta) B_n^{-j}(\theta) \right]^{-1} dG_n(u)}.$$

Since $n_j = \mathbb{P}_n 1\{\delta = j\}$ and $G_{nj}(u) = n_j^{-1} \mathbb{P}_n 1\{Y \le u, \delta = j\}$, $\hat{A}_\theta$ is a function of the empirical distribution $\mathbb{P}_n$. Hence $\hat{A}_\theta^\circ$ is obtained by replacing $\mathbb{P}_n$ with $\mathbb{P}_n^\circ$ in the definition of $\hat{A}_\theta$ and its components (eg., $\lambda_{nj}$, for $j = 0, 1$, $B_n$, etc.). Dixon, Kosorok and Lee (2005) verify the conditions of Corollary 20.3 for this setting, and thus the piggyback bootstrap is valid, provided appropriate draws $\theta_n$ can be obtained.

Dixon, Kosorok and Lee (2005) also establish all of the conditions of Theorem 19.6 for the partial-profile likelihood $p\ell_n(\theta)$ except for Condition (19.17). Nevertheless, this condition probably holds, although we will not verify it here. Alternatively, one can estimate the limiting covariance matrix for $\sqrt{n}(\hat{\theta}_n - \theta_0)$ using Corollary 19.4 which is easy to implement in this instance since $\theta$ is a scalar. Call the resulting covariance estimator $\hat{V}$. Then one can generate $\theta_n = \hat{\theta}_n + Z_n \sqrt{\hat{V}/n}$, where the $Z_n$ is a standard normal deviate. Hence $(\theta_n, \hat{A}^\circ_{\theta_n})$ will be a jointly valid deviate that can be used for joint inference on $(\theta, A)$.

### *20.2.3 Other Methods*

There are a number of other methods that can be used for inference in the setting of this chapter, including fully Bayesian approaches, although

none of them are as applicable in generality as the bootstrap or piggyback bootstrap, as far as we are aware. Nevertheless, there are a few worthwhile methods of inference that apply to specific settings.

For the Cox model under right censoring, Lin, Fleming and Wei (1994) developed an influence function approach for inference for the baseline hazard. The basic idea—which is broadly applicable when the influence function is a smooth, closed form function of the parameter—is to estimate the full influence function $\tilde{\psi}(X)$ for the joint parameters at observation $X$ by plugging in the maximum likelihood estimators in place of the true parameter values. Let the resulting estimated influence function applied to observation $X$ be denoted $\check{\psi}(X)$. Let $G_1, \ldots, G_n$ be i.i.d. standard normals, and define $\hat{H}_n \equiv n^{-1/2} \sum_{i=1}^n G_i \check{\psi}(X_i)$, and let $H$ be the tight, mean zero Gaussian process limiting distribution of $\sqrt{n}(\hat{\psi}_n - \psi_0) = n^{-1/2} \sum_{i=1}^n \tilde{\psi}(X_i) + o_{P_0}(1)$, where the error term is uniform. Then, provided $\check{\psi}(X)$ and $\tilde{\psi}(X)$ live in a Donsker class with probability approaching 1 as $n \to \infty$ and provided $P \left\| \check{\psi}(X) - \tilde{\psi}(X) \right\|_\infty^2 \xrightarrow{\mathrm{P}} 0$, it is easy to see that $\hat{H}_n \underset{G}{\rightsquigarrow} H$. Although it is computationally easy, this general approach is somewhat narrowly applicable because most influence functions do not have a sufficiently simple, closed form.

Another approach for the Cox model under right censoring, proposed by Kim and Lee (2003), is to put certain priors on the full likelihood and generate realizations of the posterior distribution. Kim and Lee derive an elegant Bernstein-von Mises type result which shows that the resulting posterior, conditional on the data, converges weakly to the correct joint limiting distribution of the joint maximum likelihood estimators. This approach is only applicable for models with very special structure, but it is quite computationally efficient when it can be applied.

Hunter and Lange (2002) develop an interesting accelerated bootstrap procedure for the proportional odds model under right censoring which facilitates faster maximization, resulting in a faster computation of bootstrap realizations. Because the method relies on the special structure of the model, it is unclear whether the procedure can be more broadly generalized. There are probably a number of additional inferential methods that apply to various specific but related problems in maximum likelihood inference that we have failed to mention. Our omission is unintentional, and we invite the interested reader to search out these methods more completely than we have. Nevertheless, the weighted bootstrap and piggyback bootstrap approaches we have discussed are among the most broadly applicable and computationally feasible methods available.

## 20.3 Exercises

**20.3.1.** Show that Condition (20.4) implies that $\Psi$ is Fréchet-differentiable in $\ell^\infty(\mathcal{C}_p)$.

**20.3.2.** Show that if Conditions (20.2)–(20.4) hold for any $p > 0$, then they hold for all $0 < p < \infty$. Hint: Use the linearity of the operators involved.

**20.3.3.** Prove Corollary 20.1.

**20.3.4.** Prove that the joint maximum likelihood estimator $(\hat{\theta}_n, \hat{A}_n)$ for the proportional odds model satisfies the conditions of Corollary 20.1. Thus $(\hat{\theta}_n, \hat{A}_n)$ is asymptotically efficient.

**20.3.5.** Show that (20.5) forces $a = 0$ and $h = 0$.

**20.3.6.** Show that Conditions (20.2)–(20.4) hold for $p = 1$ in the biased sampling example given at the end of Section 20.1.

**20.3.7.** Show that $\sigma_{22} : \mathcal{H} \mapsto \mathcal{H}$ for the proportional odds regression model for right censored data is continuously invertible and onto. Note that $\sigma_{22}$ corresponds to $\sigma^{22}_{\theta_0}$ as defined in (15.20).

## 20.4 Notes

The general semiparametric model structure that emerges in Section 20.1 after the proof of Corollary 3.2 was inspired and informed by the general model framework utilized in Dixon, Kosorok and Lee (2005). The biased sampling example was also studied in Dixon, Kosorok and Lee, and Corollary 20.3 is essentially Theorem 2.1 of Dixon, Kosorok and Lee, although the respective proofs are quite different.

# 21
# Semiparametric M-Estimation

The purpose of this chapter is to extend the M-estimation results of Chapter 14 to semiparametric M-estimation, where there is both a Euclidean parameter of interest $\theta$ and a nuisance parameter $\eta$. Obviously, the semiparametric maximum likelihood estimators we have been discussing in the last several chapters are important examples of semiparametric M-estimators, where the objective function is an empirical likelihood. However, there are numerous other examples of semiparametric M-estimators, including estimators obtained from misspecified semiparametric likelihoods, least-squares, least-absolute deviation, and penalized maximum likelihood (which we discussed some in Sections 4.5 and 15.1 and elsewhere). In this chapter, we will try to provide general results on estimation and inference for semiparametric M-estimators, along with several illustrative examples. The material for this chapter is adapted from Ma and Kosorok (2005b).

As with Chapter 19, the observed data of interest consist of $n$ i.i.d. observations $X_1 \ldots X_n$ drawn from a semiparametric model $\{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$, where $\Theta$ is an open subset of $\mathbb{R}^k$ and $H$ is a possibly infinite-dimensional set. The parameter of interest is $\theta$ and $\psi = (\theta, \eta)$ will be used to denote the joint parameter. The criterion functions we will focus on will primarily be of the form $\mathbb{P}_n m_\psi(X)$, for some class of measurable functions $\{m_\psi\}$. The most widely used M-estimators of this kind include maximum likelihood (MLE), misspecified MLE, ordinary least squares (OLS), and least absolute deviation estimators. A useful general theorem for investigating the asymptotic behavior of M-estimators for the Euclidean parameter in semiparametric models, which we will build upon in this chapter, is given in Wellner, Zhang and Liu (2002).

An alternative approach to obtaining the limit distribution of M-estimators is to view them as Z-estimators based on estimating equations obtained by differentiating the involved objective functions. We have used this approach several times for maximum likelihood estimation, including in Section 15.2 for the proportional odds model under right censoring. An advantage of this approach is that the Z-estimator theory of Chapter 13 can be utilized. The Z-estimator approach is particularly useful when the joint estimator $\hat{\psi}_n = (\hat{\theta}_n, \hat{\eta}_n)$ has convergence rate $\sqrt{n}$. Unfortunately, this theory usually does not apply when some of the parameters cannot be estimated at the $\sqrt{n}$ rate.

Penalized M-estimation, as illustrated previously for the partly linear logistic regression model (see Sections 4.5 and 15.1), provides a flexible alternative to ordinary M-estimation. Other examples of penalized M-estimators for semiparametric models include the penalized MLE for generalized partial linear models of Mammen and van de Geer (1997), the penalized MLE for the Cox model with interval censored data of Cai and Betensky (2003), and the penalized MLE for transformation models with current status data of Ma and Kosorok (2005a).

The main innovation of Ma and Kosorok (2005b) is the development of bootstrap theory for M-estimation that does not require subsampling (as in Politis and Romano, 1994), yet permits the nuisance parameter to be slower than $\sqrt{n}$-consistent. It is interesting to note that this development occur empirical process bootstrap (see Mason and Newton, 1992, Barbe and Bertail, 1995, Giné, 1996, and Wellner and Zhan, 1996). The specific bootstrap studied in Ma and Kosorok (2005b) is the weighted bootstrap which, as mentioned in previous chapters, consists of i.i.d. positive random weights applied to each observation. The reason for using the weighted bootstrap is that the i.i.d. behavior of the weights makes many of the proofs easier. While it is possible that our results also hold for the nonparametric bootstrap, such a determination is beyond the scope of this chapter and appears to be quite difficult.

The overarching goal of this chapter is to develop a paradigm for semiparametric M-estimators that facilitates a conceptual link between weak convergence and validation of the bootstrap, even when the nuisance parameter may not be $\sqrt{n}$-estimable. To this end, the first main result we present is a general theory for weak convergence, summarized in Theorem 21.1 below. The theorem mentioned earlier in Wellner, Zhang and Liu (2002) is a corollary of our Theorem 21.1. A great variety of models can be analyzed by the resulting unified approach. The second main result is a validation of the use of the weighted bootstrap as a universal inference tool for the parametric component, even when the nuisance parameters cannot be estimated at the $\sqrt{n}$ convergence rate and/or when the usual inferences based on the likelihood principle do not apply, as happens for example in the penalized M-estimation settings. We also show how these two main

results can be linked through a careful analysis of the entropy numbers of the objective functions.

The layout of the chapter is as follows. In Section 21.1, after first presenting some motivating examples, we develop $\sqrt{n}$-consistency and asymptotic normality results for the estimators of the Euclidean parameters. In Section 21.2, weighted M-estimators with independent weights are studied and these results are used to establish validity of the weighted bootstrap as an inference tool for the Euclidean parameter. Control of the modulus of continuity of the weighted empirical processes plays an important role in our approach. This can be viewed as an extension of the principle utilized in the rate of convergence theorem (Theorem 14.4) of Section 14.3. The connection mentioned above with the entropy numbers of the involved objective functions is described in detail in Section 21.3. In Section 21.4, we study in greater detail the examples of Section 21.1 and verify consistency and bootstrap validity using the proposed techniques. The chapter concludes in Section 21.5 with an extension to penalized M-estimation.

## 21.1 Semiparametric M-estimators

The purpose of this section is to provide a general framework for estimation and weak convergence of M-estimators for the parametric component $\theta$ in a semiparametric model. We first present several motivating examples. Then we present a general scheme for semiparametric M-estimation. We then discuss consistency and rate of convergence, followed by two different approaches to establishing asymptotic normality of the estimator. As with previous chapters, we let $P$ denote expectation under the true distribution.

### *21.1.1 Motivating Examples*

Although only three examples will be given here, they stand as archetypes for a great variety of models that can be studied in a similar manner. In the following sections, derivatives will be denoted with superscript "()". For example, $h^{(2)}$ refers to the second order derivative of the function $h$. Notice that each of these examples departs in some way from the usual semiparametric MLE paradigm.

*The Cox model with current status data (Example 1).*

This is an augmentation of the example presented in Section 19.2.3. Let $\theta$ be the regression coefficient and $\Lambda$ the baseline integrated hazard function. The MLE approach to inference for this model was discussed at the end of both Sections 19.3.1 and 19.3.2. As an alternative estimation approach, $(\theta, \Lambda)$ can also be estimated by OLS:

$$(\hat{\theta}_n, \hat{\Lambda}_n) = \text{argmin}\, \mathbb{P}_n \left(1 - \delta_i - \exp(-e^{\theta' Z_i} \Lambda(t_i)\right)^2 .$$

In this model, the nuisance parameter $\Lambda$ cannot be estimated at the $\sqrt{n}$ rate, but is estimable at the $n^{1/3}$ rate, as discussed previously (see also Groeneboom and Wellner, 1992).

*Binary regression under misspecified link function (Example 2).*

This is a modification of the partly linear logistic regression model introduced in Chapter 1 and discussed in Sections 4.5 and 15.1. Suppose that we observe an i.i.d. random sample $(Y_1, Z_1, U_1), \ldots, (Y_n, Z_n, U_n)$ consisting of a binary outcome $Y$, a $k-$dimensional covariate $Z$, and a one-dimensional continuous covariate $U \in [0,1]$, following the additive model

$$P_{\theta,h}(Y = 1 | Z = z, U = u) = \phi(\theta' z + h(u)),$$

where h is a smooth function belonging to

$$\mathbb{H} = \left\{ h : [0,1] \mapsto [-1,1], \int_0^1 \left(h^{(s)}(u)\right)^2 du \leq K \right\}, \tag{21.1}$$

for a fixed and known $K \in (0, \infty)$ and an integer $s \geq 1$, and where $\phi : \mathbb{R} \mapsto [0,1]$ is a known continuously differentiable monotone function. The choices $\phi(t) = 1/(1 + e^{-t})$ and $\phi = \Phi$ (the cumulative normal distribution function) correspond to the logit model and probit models, respectively. The maximum likelihood estimator $(\hat{\theta}_n, \hat{h}_n)$ maximizes the (conditional) log-likelihood function

$$\begin{aligned} \ell_n(\theta, h)(X) \quad = \quad & \mathbb{P}_n \left[ Y \log(\phi(\theta' Z + h(U))) \right. \\ & \left. + (1 - Y) \log(1 - \phi(\theta' Z + h(U))) \right], \end{aligned} \tag{21.2}$$

where $X = (Y, Z, U)$. Here, we investigate the estimation of $(\theta, h)$ under misspecification of $\phi$. Under model misspecification, the usual theory for semiparametric MLEs, as discussed in Chapter 19, does not apply.

The condition $\int_0^1 \left(h^{(s)}(u)\right)^2 du \leq K$ in (21.1) can be relaxed as follows. Instead of maximizing the log-likelihood as in (21.2), we can take $(\hat{\theta}_n, \hat{h}_n)$ to be the maximizer of the penalized log-likelihood $\ell_n(\theta, h) - \lambda_n^2 J^2(h)$, as done in Chapter 1, where $J$ is as defined in Chapter 1, and where $\lambda_n$ is a data-driven smoothing parameter. Then we only need to assume $\int_0^1 \left(h^{(s)}(u)\right)^2 du < \infty$. The application of penalization to the misspecified model will be examined more closely in Section 21.5.

*Mixture models (Example 3).*

Suppose that an observation $X$ has a conditional density $p_\theta(x|z)$ given an unobservable variable $Z = z$, where $p_\theta$ is known up to the Euclidean parameter $\theta$. If the unobservable $Z$ possesses an unknown distribution $\eta$, then the

observation $X$ has the following mixture density $p_{\theta,\eta}(x) = \int p_\theta(x|z)d\eta(z)$. The maximum likelihood estimator $(\hat{\theta}_n, \hat{\eta}_n)$ maximizes the log-likelihood function $(\theta, \eta) \mapsto \ell_n(\theta, \eta) \equiv \mathbb{P}_n \log(p_{\theta,\eta}(X))$.

Examples of mixture models include frailty models, errors-in-variable models in which the errors are modeled by a Gaussian distribution, and scale mixture models over symmetric densities. For a detailed discussion of the MLE for semiparametric mixture models, see van der Vaart (1996) and van de Geer (2000).

### *21.1.2 General Scheme for Semiparametric M-Estimators*

Consider a semiparametric statistical model $P_{\theta,\eta}(X)$, with $n$ i.i.d. observations $X_1 \ldots X_n$ drawn from $P_{\theta,\eta}$, where $\theta \in \mathbb{R}^k$ and $\eta \in H$. Assume that the infinite dimensional space $H$ has norm $\|\cdot\|$, and the true unknown parameter is $(\theta_0, \eta_0)$. An M-estimator $(\hat{\theta}_n, \hat{\eta}_n)$ of $(\theta, \eta)$ has the form

$$(\hat{\theta}_n, \hat{\eta}_n) = \operatorname{argmax} \mathbb{P}_n m_{\theta,\eta}(X), \tag{21.3}$$

where $m$ is a known, measurable function. All of the following results of this chapter will hold, with only minor modifications, if "argmax" in Equation (21.3) is replaced by "argmin". For simplicity, we assume the limit criterion function $Pm_\psi$, where $\psi = (\theta, \eta)$, has a unique and "well-separated" point of maximum $\psi_0$, i.e., $Pm_{\psi_0} > \sup_{\psi \notin G} Pm_\psi$ for every open set $G$ that contains $\psi_0$.

Analysis of the asymptotic behavior of M-estimators can be split into three main steps: (1) establishing consistency; (2) establishing a rate of convergence; and (3) deriving the limiting distribution.

A typical scheme for studying general semiparametric M-estimators is as follows. First, consistency is established with the argmax theorem (Theorem 14.1) or a similar method. Second, the rate of convergence for the estimators of all parameters can then be obtained from convergence rate theorems such as Theorem 14.4. We briefly discuss in Section 21.1.3 consistency and rate of convergence results in the semiparametric M-estimation context. The asymptotic behavior of estimators of the Euclidean parameters can be studied with Theorem 21.1 or Corollary 21.2 presented in Section 21.1.4 below.

Important properties of weighted M-estimators that enable validation of the weighted bootstrap are described in Section 21.2 below. Lemmas 21.8–21.10 in Section 21.3 can be used to control the modulus of continuity of weighted M-estimators, as needed for bootstrap validation, based on modulus of continuity results for the unweighted M-estimators. The basic conclusion is that bootstrap validity of weighted M-estimators is almost an automatic consequence of modulus of continuity control for the corresponding unweighted M-estimator. The remainder of this Section is devoted to elucidating this general scheme.

### 21.1.3 Consistency and Rate of Convergence

The first steps are to establish consistency and rates of convergence for all parameters of the unweighted (original) M-estimator. General theory for these aspects was studied in detail in Chapter 14.

Consistency of M-estimators can be achieved by careful application of the argmax theorem, as discussed, for example, in Section 14.2. Application of the argmax theorem often involves certain compactness assumptions on the parameter sets along with model identifiability. In this context, it is often sufficient to verify that the class of functions $\{m_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$ is $P$-Glivenko-Cantelli. Such an approach is used in the proof of consistency for Example 1, given below in Section 21.4. More generally, establishing consistency can be quite difficult. While a streamlined approach to establishing consistency is not a primary goal of this chapter, we will present several useful tools for accomplishing this in the examples and results that follow.

The basic tool in establishing the rate of convergence for an M-estimator is control of the modulus of continuity of the empirical criterion function using entropy integrals over the parameter sets. Entropy results in Section 11.1 and in van de Geer (2000) give rate of convergence results for a large variety of models, as we will demonstrate for Examples 1–3 in Section 21.4 below.

### 21.1.4 $\sqrt{n}$ Consistency and Asymptotic Normality

In this section, we develop theory for establishing $\sqrt{n}$ consistency and asymptotic normality for the maximizer $\hat{\theta}_n$ obtained from a semiparametric objection function $m$. We present two philosophically different approaches, one based on influence functions and one based on score equations.

*An influence function approach.*

We now develop a paradigm for studying the asymptotic properties of $\hat{\theta}_n$, based on an arbitrary $m$, which parallels the efficient influence function paradigm used for MLE's (where $m$ is the log-likelihood).

For any fixed $\eta \in H$, let $\eta(t)$ be a smooth curve running through $\eta$ at $t = 0$, that is $\eta(0) = \eta$. Let $a = (\partial/\partial t)\eta(t)|_{t=0}$ be a proper tangent in the tangent set $\dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}}$ for the nuisance parameter. The concepts of tangent (or "direction of perturbation") and tangent set/space are identical in this context as in the usual semiparametric context of Section 3.2, since tangent sets come form the model and not from the method of estimation. Accordingly, the requirements of the perturbation $a$ are the usual ones, including the properties that $a \in L_2(P)$ and, when $t$ is small enough, that $\eta(t) \in H$. To simplify notation some, we will use $\mathbb{A}$ to denote $\dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}}$.

For simplicity, we denote $m(\theta, \eta; X)$ as $m(\theta, \eta)$. Set

$$m_1(\theta,\eta) = \frac{\partial}{\partial\theta} m(\theta,\eta), \text{ and } m_2(\theta,\eta)[a] = \frac{\partial}{\partial t}\bigg|_{t=0} m(\theta,\eta(t)),$$

where $a \in \mathbb{A}$. We also define

$$m_{11}(\theta,\eta) = \frac{\partial}{\partial\theta} m_1(\theta,\eta), \quad m_{12}(\theta,\eta)[a] = \frac{\partial}{\partial t}\bigg|_{t=0} m_1(\theta,\eta(t)), \text{ and}$$

$$m_{21}(\theta,\eta)[a] = \frac{\partial}{\partial\theta} m_2(\theta,\eta)[a], \quad m_{22}(\theta,\eta)[a_1][a_2] = \frac{\partial}{\partial t}\bigg|_{t=0} m_2(\theta,\eta_2(t))[a_1],$$

where $a, a_1$ and $a_2 \in \mathbb{A}$, and $(\partial/(\partial t))\eta_2(t)|_{t=0} = a_2$.

A brief review of the construction of maximum likelihood estimators is insightful here. In this special case, $m$ is a log-likelihood. Denote by $[m_2]$ the linear span of the components of $m_2$ in $L_2(P)$. Recall that the efficient score function $\tilde{m}$ is equal to the projection of the score function $m_1$ onto the orthocomplement of $[m_2]$ in $L_2(P)$. As mentioned in Section 3.2, one way of estimating $\theta$ is by solving the efficient score equations $\mathbb{P}_n \tilde{m}_{\theta,\hat{\eta}_n}(X) = 0$.

For general M-estimators, a natural extension is to construct estimating equations as follows. Define $m_2(\theta,\eta)[A] = (m_2(\theta,\eta)[a_1], \ldots, m_2(\theta,\eta)[a_k])$, where $A = (a_1, \ldots, a_k)$ and $a_1, \ldots, a_k \in \mathbb{A}$. Then $m_{12}[A_1]$ and $m_{22}[A_1][A_2]$ can be defined accordingly, for $A_1 = (a_{11}, \ldots, a_{1k})$, $A_2 = (a_{21}, \ldots, a_{2k})$ and $a_{i,j} \in \mathbb{A}$. Assume there exists an $A^* = (a_1^*, \ldots, a_k^*)$, where $\{a_i^*\} \in \mathbb{A}$, such that for any $A = (a_1, \ldots, a_k)$, $\{a_i\} \in \mathbb{A}$,

$$P\left(m_{12}(\theta_0,\eta_0)[A] - m_{22}(\theta_0,\eta_0)[A^*][A]\right) = 0. \tag{21.4}$$

Define $\tilde{m}(\theta,\eta) \equiv m_1(\theta,\eta) - m_2(\theta,\eta)[A^*]$. $\theta$ is then estimated by solving $\mathbb{P}_n \tilde{m}(\theta,\hat{\eta}_n; X) = 0$, where we substitute an estimator $\hat{\eta}_n$ for the unknown nuisance parameter. A variation of this approach is to obtain an estimator $\hat{\eta}_n(\theta)$ of $\eta$ for each given value of $\theta$ and then solve $\theta$ from

$$\mathbb{P}_n \tilde{m}(\theta, \hat{\eta}_n(\theta); X) = 0. \tag{21.5}$$

In some cases, estimators satisfying (21.5) may not exist. Hence we weaken (21.5) to the following "nearly-maximizing" condition:

$$\mathbb{P}_n \tilde{m}(\hat{\theta}_n, \hat{\eta}_n) = o_P(n^{-1/2}). \tag{21.6}$$

We next give sufficient conditions for $\hat{\theta}_n$, based on (21.6), to be $\sqrt{n}$ consistent and asymptotically normally distributed:

A1: (Consistency and rate of convergence) Assume

$$|\hat{\theta}_n - \theta_0| = o_P(1), \text{ and } \|\hat{\eta}_n - \eta_0\| = O_P(n^{-c_1}),$$

for some $c_1 > 0$, where $|\cdot|$ will be used in this chapter to denote the Euclidean norm.

A2: (Finite variance) $0 < \det(I^*) < \infty$, where det denotes the determinant of a matrix and

$$\begin{aligned} I^* &= \{P(m_{11}(\theta_0,\eta_0) - m_{21}(\theta_0,\eta_0)[A^*])\}^{-1} \\ &\quad \times P[m_1(\theta_0,\eta_0) - m_2(\theta_0,\eta_0)[A^*]]^{\otimes 2} \\ &\quad \times \{P(m_{11}(\theta_0,\eta_0) - m_{21}(\theta_0,\eta_0)[A^*])\}^{-1}, \end{aligned}$$

where superscript $\otimes 2$ denotes outer product.

A3: (Stochastic equicontinuity) For any $\delta_n \downarrow 0$ and $C > 0$,

$$\sup_{|\theta-\theta_0|\le\delta_n, \|\eta-\eta_0\|\le Cn^{-c_1}} \left|\sqrt{n}(\mathbb{P}_n - P)(\tilde{m}(\theta,\eta) - \tilde{m}(\theta_0,\eta_0))\right| = o_P(1).$$

A4: (Smoothness of the model) For some $c_2 > 1$ satisfying $c_1 c_2 > 1/2$ and for all $(\theta,\eta)$ satisfying $\{(\theta,\eta) : |\theta-\theta_0| \le \delta_n,\ \|\eta-\eta_0\| \le Cn^{-c_1}\}$,

$$\begin{aligned} &\Bigg| P\Bigg\{ (\tilde{m}(\theta,\eta) - \tilde{m}(\theta_0,\eta_0)) - \left(m_{11}(\theta_0,\eta_0) - m_{21}(\theta_0,\eta_0)[A^*]\right)(\theta-\theta_0) \\ &- \left( m_{12}(\theta_0,\eta_0)\left[\frac{\eta-\eta_0}{\|\eta-\eta_0\|}\right] - m_{22}(\theta_0,\eta_0)[A^*]\left[\frac{\eta-\eta_0}{\|\eta-\eta_0\|}\right]\right)\|\eta-\eta_0\| \Bigg\}\Bigg| \\ &\qquad = o(|\theta-\theta_0|) + O(\|\eta-\eta_0\|^{c_2}). \end{aligned}$$

THEOREM 21.1 *Suppose that $(\hat{\theta}_n, \hat{\eta}_n)$ satisfies Equation (21.6), and that Conditions A1–A4 hold, then*

$$\begin{aligned} &\sqrt{n}(\hat{\theta}_n - \theta_0) \\ &\quad = -\sqrt{n}\{P(m_{11}(\theta_0,\eta_0) - m_{21}(\theta_0,\eta_0)[A^*])\}^{-1} \\ &\qquad \times \mathbb{P}_n(m_1(\theta_0,\eta_0) - m_2(\theta_0,\eta_0)[A^*]) + o_P(1). \end{aligned}$$

*Hence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and variance $I^*$.*

PROOF. Combining Condition A1 and A3, we have

$$\sqrt{n}(\mathbb{P}_n - P)\tilde{m}(\hat{\theta}_n,\hat{\eta}_n) = \sqrt{n}(\mathbb{P}_n - P)\tilde{m}(\theta_0,\eta_0) + o_P(1). \tag{21.7}$$

Considering Equation (21.6), we can further simplify Equation (21.7) to

$$\sqrt{n}P(\tilde{m}(\hat{\theta}_n,\hat{\eta}_n) - \tilde{m}(\theta_0,\eta_0)) = -\sqrt{n}\mathbb{P}_n\tilde{m}(\theta_0,\eta_0) + o_P(1). \tag{21.8}$$

Considering Conditions A1 and A4, we can expand the left side of Equation (21.8) to obtain

(21.9)

$$\begin{aligned} & P(m_{11}(\theta_0,\eta_0) - m_{21}(\theta_0,\eta_0)[A^*])(\hat{\theta}_n - \theta_0) \\ & \qquad + P\left( m_{12}(\theta_0,\eta_0)\left[\frac{\hat{\eta}_n - \eta_0}{\|\hat{\eta}_n - \eta_0\|}\right] - m_{22}(\theta_0,\eta_0)[A^*]\left[\frac{\hat{\eta}_n - \eta_0}{\|\hat{\eta}_n - \eta_0\|}\right]\right) \\ & \qquad \times \|\hat{\eta}_n - \eta_0\| \\ = \;& o_P(|\hat{\theta}_n - \theta_0|) + O_P(\|\hat{\eta}_n - \eta_0\|^{c_2}) - \mathbb{P}_n \tilde{m}(\theta_0,\eta_0) + o_P(n^{-1/2}). \end{aligned}$$

Equation (21.9), Condition A1 and Condition A2 together give us

$$\begin{aligned} & \sqrt{n} P(m_{11}(\theta_0,\eta_0) - m_{21}(\theta_0,\eta_0)[A^*])(\hat{\theta}_n - \theta_0) \\ & \qquad = \; -\sqrt{n}\mathbb{P}_n \tilde{m}(\theta_0,\eta_0) + o_P(1), \end{aligned}$$

and Theorem 21.1 follows.□

Condition A1 can be quite difficult to establish. Some of these challenges were discussed in Section 21.1.3 above. In some cases, establishing A1 can be harder than establishing all of the remaining conditions of Theorem 21.1 combined. Fortunately, there are various techniques for attacking the problem, and we will outline some of them in the examples considered in Section 21.4 below. Condition A2 corresponds to the nonsingular information condition for the MLE. The asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is degenerate if this condition is not satisfied. For the case of the MLE, $I^*$ can be further simplified to $-\{P(m_{11}(\theta_0,\eta_0) - m_{21}(\theta_0,\eta_0)[A^*])\}^{-1}$.

Conditions A3 and A4 are perhaps less transparent than Conditions A1 and A2, but a number of techniques are available for verification. Condition A3 can be verified via entropy calculations and certain maximal inequalities, for example, as discussed in Chapter 14 and as demonstrated in Huang (1996) and in van der Vaart (1996). One relatively simple, sufficient condition is for the class of functions $\{\tilde{m}(\theta,\eta) : |\theta - \theta_0| \le \epsilon_1, \|\eta - \eta_0\| \le \epsilon_2\}$ to be Donsker for some $\epsilon_1, \epsilon_2 > 0$ and that $P\left(\tilde{m}(\theta,\eta) - \tilde{m}(\theta_0,\eta_0)\right)^2 \to 0$ as both $|\theta - \theta_0| \to 0$ and $\|\eta - \eta_0\| \to 0$.

Condition A4 can be checked via Taylor expansion techniques for functionals. Roughly speaking, if the model is differentiable in the ordinary sense when we replace the nonparametric parameter with a Euclidean parameter, then often the right kind of smoothness for the infinite dimensional parameter $\eta$ also holds. For example, if all third derivatives with respect to both $\theta$ and $\eta$ of $P\tilde{m}(\theta,\eta)$ are bounded in a neighborhood of $(\theta_0,\eta_0)$, then the expression in Condition A4 holds for $c_2 = 2$. Hence Condition A4 will hold provided $c_1 > 1/4$. Fortunately, $c_1 > 1/4$ for many settings, including all of the examples considered in Section 21.4 below.

*A score equation approach.*

Another way of looking at M-estimators, which is perhaps more intuitive, is as follows. By definition, M-estimators maximize an objective function, i.e.,

$$(\hat{\theta}_n, \hat{\eta}_n) = \operatorname{argmax} \mathbb{P}_n m(\theta, \eta; X). \tag{21.10}$$

From Equation (21.10), we have

$$\mathbb{P}_n m_1(\hat{\theta}_n, \hat{\eta}_n) = 0, \quad \text{and} \quad \mathbb{P}_n m_2(\hat{\theta}_n, \hat{\eta}_n)[a] = 0, \tag{21.11}$$

where $a$ runs over $\mathbb{A}$. Viewed in this fashion, M-estimators are also Z-estimators. Hence the theory of Chapter 13 can be utilized. We can relax (21.11) to the following "nearly-maximizing" conditions:

$$\begin{aligned} \mathbb{P}_n m_1(\hat{\theta}_n, \hat{\eta}_n) &= o_P(n^{-1/2}), \text{ and} \\ \mathbb{P}_n m_2(\hat{\theta}_n, \hat{\eta}_n)[a] &= o_P(n^{-1/2}), \text{ for all } a \in \mathbb{A}. \end{aligned} \tag{21.12}$$

The following corollary provides sufficient conditions under which estimators satisfying (21.12) and Conditions A1–A2 in Theorem 21.1 will have the same properties obtained in Theorem 21.1. Before giving the result, we articulate two substitute conditions for A3–A4 which are needed in this setting:

B3: (Stochastic equicontinuity) For any $\delta_n \downarrow 0$ and $C > 0$,

$$\begin{aligned} &\sup_{|\theta-\theta_0|\le\delta_n, \|\eta-\eta_0\|\le Cn^{-c_1}} \left|\sqrt{n}(\mathbb{P}_n - P)\left(m_1(\theta,\eta) - m_1(\theta_0,\eta_0)\right)\right| \\ &\quad = \; o_P(1), \text{ and} \\ &\sup_{|\theta-\theta_0|\le\delta_n, \|\eta-\eta_0\|\le Cn^{-c_1}} \left|\sqrt{n}(\mathbb{P}_n - P)\left(m_2(\theta,\eta) - m_2(\theta_0,\eta_0)\right)[A^*]\right| \end{aligned}$$

$= o_P(1)$, where $c_1$ is as in Condition A1.

B4: (Smoothness of the model) For some $c_2 > 1$ satisfying $c_1 c_2 > 1/2$, where $c_1$ is given in Condition A1, and for all $(\theta, \eta)$ belonging to $\{(\theta,\eta) : |\theta - \theta_0| \le \delta_n,\ \|\eta - \eta_0\| \le Cn^{-c_1}\}$,

$$\begin{aligned} \Bigg| P \Bigg\{ & m_1(\theta,\eta) - m_1(\theta_0,\eta_0) - m_{11}(\theta_0,\eta_0)(\theta-\theta_0) \\ & - m_{12}(\theta_0,\eta_0)\left[\frac{\eta-\eta_0}{\|\eta-\eta_0\|}\right] \times \|\eta-\eta_0\| \Bigg\} \Bigg| \\ & = o(|\theta-\theta_0|) + O(\|\eta-\eta_0\|^{c_2}), \end{aligned}$$

and

$$\begin{aligned} \Bigg| P \Bigg\{ & m_2(\theta,\eta)[A^*] - m_2(\theta_0,\eta_0)[A^*] - m_{21}(\theta_0,\eta_0)[A^*](\theta-\theta_0) \\ & - m_{22}(\theta_0,\eta_0)[A^*]\left[\frac{\eta-\eta_0}{\|\eta-\eta_0\|}\right] \times \|\eta-\eta_0\| \Bigg\} \Bigg| \\ & = o(|\theta-\theta_0|) + O(\|\eta-\eta_0\|^{c_2}). \end{aligned}$$

COROLLARY 21.2 *Suppose that the estimator $(\hat{\theta}_n, \hat{\eta}_n)$ satisfies Equation (21.12), and Conditions A1, A2, B3 and B4 all hold. Then the results of Theorem 21.1 hold for $\hat{\theta}_n$.*

Setting $A = A^*$ and combining the two equations in (21.12), we obtain (21.6). Subtracting the two equations in Conditions B3 and B4 in Corollary 21.2, we can obtain Conditions A3 and A4 in Theorem 21.1, respectively. Thus the conditions in Corollary 21.2 are stronger than their counterparts in Theorem 21.1. However, Corollary 21.2 is sometimes easier to understand and implement. We also note that simpler sufficient conditions for B3 and B4 can be developed along the lines of those developed above for Conditions A3 and A4.

## 21.2 Weighted M-Estimators and the Weighted Bootstrap

We now investigate inference for $\theta$. For the parametric MLE, the most widely used inference techniques are based on the likelihood. For general semiparametric M-estimation, a natural thought is to mimic the approach for profile log-likelihood expansion studied in Chapter 19. However, what we would then obtain is

$$\begin{aligned}\mathbb{P}_n m(\tilde{\theta}_n) &= \mathbb{P}_n m(\theta_0) + (\tilde{\theta}_n - \theta_0)' \sum_{i=1}^{n} m_1(\theta_0)(X_i) \\ &\quad - \frac{1}{2} n(\tilde{\theta}_n - \theta_0)' P \{m_{11} - m_{21}[A^*]\} (\tilde{\theta}_n - \theta_0) \\ &\quad + o_P(1 + \sqrt{n}|\hat{\theta}_n - \theta_0|)^2,\end{aligned}$$

for any sequence $\tilde{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$. Unfortunately, the curvature in the quadratic term does not correspond to the inverse of $I^*$. This makes inference based on the likelihood ratio impractical, and we will not pursue this approach further in this chapter.

In contrast, the weighted bootstrap—which can be expressed in this setting as a weighted M-estimator—appears to be an effective and nearly universal inference tool for semiparametric M-estimation. The goal of this section is to verify that this holds true in surprising generality. We first study the unconditional behavior of weighted M-estimators and then use these results to establish conditional asymptotic validity of the weighted bootstrap.

Consider $n$ i.i.d. observations $X_1, \ldots, X_n$ drawn from the true distribution $P$. Denote $\xi_i$, $i = 1, .., n$, as $n$ i.i.d. positive random weights, satisfying $\mathrm{E}(\xi) = 1$ and $0 \leq \mathrm{var}(\xi) = v_0 < \infty$ and which are independent of the data $\mathcal{X}_n = \sigma(X_1, \ldots, X_n)$, where $\sigma(U)$ denotes the smallest $\sigma$-algebra for which

$U$ is measurable. Note that the weights are very similar to the ones introduced for the weighted bootstrap in Section 2.2.3, except that $\mathrm{E}(\xi) = 1$ and we do not require $\|\xi\|_{2,1} < \infty$ nor do we divide the weights by $\bar{\xi}$. The weighted M-estimator $(\hat{\theta}_n^\circ, \hat{\eta}_n^\circ)$ satisfies

$$(\hat{\theta}_n^\circ, \hat{\eta}_n^\circ) = \operatorname{argmax} \mathbb{P}_n \{\xi m(\theta, \eta; X)\}. \tag{21.13}$$

Asymptotic properties of weighted M-estimators will be studied in a fashion parallel to those of ordinary M-estimators.

Since we assume the random weights are independent of $\mathcal{X}_n$, the consistency and rate of convergence for the estimators of all parameters can be established using Theorems 2.12 and 14.4, respectively, after minor modifications. For completeness, we now present these slightly modified versions, Corollaries 21.3 and 21.4 below. Since these are somewhat obvious generalizations of Theorems 2.12 and 14.4, we save their proofs as exercises (see Exercise 21.6.1). In both cases, one trivially gets the results for the unweighted empirical measure $\mathbb{P}_n$ by assuming the weights $\xi_1, \ldots, \xi_n$ in $\mathbb{P}_n^\circ$ are 1 almost surely. We denote $\psi = (\theta, \eta)$ for simplicity, where $\psi \in \Gamma$, and assume there exists a semimetric $d$ making $(\Gamma, d)$ into a semimetric space. We also enlarge $P$ to be the product measure of the true distribution and the distribution of the independent weights.

Corollary 21.3 *Consistency of weighted M-estimators: Let $m$ be a measurable function indexed by $(\Gamma, d)$ such that $Pm : \Gamma \mapsto \mathbb{R}$ is deterministic. Also let $\mathbb{P}_n^\circ \equiv n^{-1} \sum_{i=1}^n \xi_i \delta_{X_i}$, where $\xi_1, \ldots, \xi_n$ are i.i.d. positive weights independent of the data $X_1, \ldots, X_n$.*

*(i) Suppose that $\|(\mathbb{P}_n^\circ - P)m\|_\Gamma \to 0$ in outer probability and there exists a point $\psi_0$ such that $Pm(\psi_0) > \sup_{\psi \notin G} Pm(\psi)$, for every open set $G$ that contains $\psi_0$. Then any sequence $\hat{\psi}_n^\circ$, such that $\mathbb{P}_n^\circ m(\hat{\psi}_n^\circ) \geq \sup_\psi \mathbb{P}_n^\circ m(\psi) - o_P(1)$, satisfies $\hat{\psi}_n^\circ \to \psi_0$ in outer probability.*

*(ii) Suppose that $\|(\mathbb{P}_n^\circ - P)m\|_K \to 0$ in outer probability for every compact $K \subset \Gamma$ and that the map $\psi \mapsto Pm(\psi)$ is upper semicontinuous with a unique maximizer at $\psi_0$. Then the same conclusion given in (i) is true provided that the sequence $\hat{\psi}_n^\circ$ is uniformly tight.*

Corollary 21.4 *Rate of convergence of weighted M-estimators: Let $m$ and $\mathbb{P}_n^\circ$ be as defined in Corollary 21.3 above. Assume also that for every $\psi$ in a neighborhood of $\psi_0$, $P(m(\psi) - m(\psi_0)) \lesssim -d^2(\psi, \psi_0)$. Suppose also that, for every $n$ and sufficiently small $\delta$, the centered process $(\mathbb{P}_n^\circ - P)(m(\psi) - m(\psi_0))$ satisfies*

$$\mathrm{E}^* \sup_{d(\psi, \psi_0) < \delta} |(\mathbb{P}_n^\circ - P)(m(\psi) - m(\psi_0))| \lesssim \frac{\phi_n(\delta)}{\sqrt{n}},$$

*for a function $\phi_n$ such that $\delta \mapsto \phi_n(\delta)/\delta^c$ is decreasing for some $c < 2$. Let $r_n^2 \phi_n(1/r_n) \leq \sqrt{n}$, for every $n$. If the sequence $\hat{\psi}_n^\circ$ satisfies*

$$\mathbb{P}_n^\circ m(\hat{\psi}_n^\circ) \geq \mathbb{P}_n^\circ m(\psi_0) - o_P(r_n^{-2})$$

*and converges in outer probability to* $\psi_0$*, then* $r_n d(\hat{\psi}_n^\circ, \psi_0) = O_P(1)$.

Assume hereafter that the estimator $(\hat{\theta}_n^\circ, \hat{\eta}_n^\circ)$ satisfies

$$\mathbb{P}_n^\circ \tilde{m}(\hat{\theta}_n^\circ, \hat{\eta}_n^\circ) = \mathbb{P}_n\{\xi \tilde{m}(\hat{\theta}_n^\circ, \hat{\eta}_n^\circ)\} = o_P(n^{-1/2}). \tag{21.14}$$

We now investigate the unconditional limiting distribution of $\hat{\theta}_n^\circ$:

COROLLARY 21.5 *Replace all* $\tilde{m}$ *in Theorem 21.1 with* $\xi\tilde{m}$*. Suppose that the estimator* $(\hat{\theta}_n^\circ, \hat{\eta}_n^\circ)$ *satisfies Equation (21.14) and Conditions A1–A4 in Theorem 21.1, then*

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n^\circ - \theta_0) &= -\sqrt{n}\,\{P(m_{11}(\theta_0, \eta_0) - m_{21}(\theta_0, \eta_0)[A^*])\}^{-1} \\ &\quad \times \mathbb{P}_n^\circ(m_1(\theta_0, \eta_0) - m_2(\theta_0, \eta_0)[A^*]) + o_P(1).\end{aligned}$$

*Thus* $\sqrt{n}(\hat{\theta}_n^\circ - \theta_0)$ *is asymptotically normally distributed with variance* $(1 + v_0)I^*$*, where* $I^*$ *is as defined in Condition A2.*

We can also obtain results for weighted M-estimators similar to those in Corollary 21.2:

COROLLARY 21.6 *Suppose the estimator* $(\hat{\theta}_n^\circ, \hat{\eta}_n^\circ)$ *satisfies*

$$\begin{aligned}\mathbb{P}_n^\circ m_1(\hat{\theta}_n^\circ, \hat{\eta}_n^\circ) &= o_P(n^{-1/2}), \text{ and} \\ \mathbb{P}_n^\circ m_2(\hat{\theta}_n^\circ, \hat{\eta}_n^\circ)[a] &= o_P(n^{-1/2}),\end{aligned} \tag{21.15}$$

*for any* $a \in \mathbb{A}$*. If we replace* $m_1$ *and* $m_2$ *with* $\xi m_1$ *and* $\xi m_2$*, respectively, then for estimators satisfying Equation (21.15) and all of the conditions of Corollary 21.2, the conclusions of Corollary 21.5 hold.*

The proofs of Corollaries 21.5 and 21.6 are direct extensions of the proofs of Theorem 21.1 and Corollary 21.2, respectively, and we save the details as an exercise (see Exercise 20.6.2).

The above results can be used to justify the use of the weighted bootstrap for general M-estimators.

THEOREM 21.7 *Suppose the M-estimator* $\hat{\theta}_n$ *and the weighted M-estimator* $\hat{\theta}_n^\circ$ *satisfy:*

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \theta_0) &= \tilde{I}_0^{-1}\sqrt{n}\mathbb{P}_n\tilde{m} + o_P(1), \quad \text{and} \\ \sqrt{n}(\hat{\theta}_n^\circ - \theta_0) &= \tilde{I}_0^{-1}\sqrt{n}\mathbb{P}_n^\circ\tilde{m} + o_P(1).\end{aligned}$$

*Assume also that the conclusions of Theorem 21.1 and Corollary 21.5 hold. Then we have* $\sqrt{n}(\hat{\theta}_n^\circ - \hat{\theta}_n) = \tilde{I}_0^{-1}\sqrt{n}(\mathbb{P}_n^\circ - \mathbb{P}_n)\tilde{m} + o_P(1)$*. Since* $\mathrm{E}(\xi) = 1$ *and* $\xi$ *is independent of* $\mathcal{X}_n$*,* $\sqrt{n/v_0}(\hat{\theta}_n^\circ - \hat{\theta}_n) \overset{\mathrm{P}}{\underset{\xi}{\rightsquigarrow}} Z_0$*, where* $\overset{\mathrm{P}}{\underset{\xi}{\rightsquigarrow}}$ *denotes conditional convergence given the data* $\mathcal{X}_n$ *and* $Z_0$ *is mean zero Gaussian with covariance* $I^*$.

Thus the weighted bootstrap is asymptotically valid for inference on $\hat{\theta}_n$. Another widely used inference tool is the nonparametric bootstrap based on multinomial weights, as discussed in Chapter 10. Consistency of the nonparametric bootstrap estimators can be established along the lines of Part (ii) of Theorems 2.6 and 2.7. However, convergence rate and asymptotic normality results (such as those in Corollary 21.5 above) are quite difficult to establish for the nonparametric bootstrap, especially for models with parameters not estimable at the $\sqrt{n}$ rate.

For the weighted bootstrap, once the asymptotic properties for the ordinary semiparametric M-estimators are established along the lines of Section 21.1, the weighted bootstrap can be verified almost automatically. The entropy control results in Section 21.3 below play an important role in connecting properties of ordinary M-estimators with the validity of the weighted bootstrap. The ease of validating the weighted bootstrap will be illustrated with the examples studied in Section 21.4.

## 21.3 Entropy Control

The asymptotic behavior of M-estimators is often closely related to certain entropy integrals, usually of the set $\{\tilde{m}(\theta, \eta) : \theta \in \Theta, \eta \in H\}$. For weighted M-estimators, the functional sets of interest are composed of functions multiplied by independent weights. The implication of the results in this section is that, in many situations, both the weighted and unweighted M-estimators can be controlled by the same entropy integral bounds. Practically, this means that the corresponding rates of convergence will also be the same. The following three lemmas are essentially generalizations of Theorems 8.4, 11.2 and 11.3, respectively. We encourage the reader to briefly review the measurability and entropy definitions and results of Chapters 8 and 9. Recall also the definitions of $J^*(\delta, \mathcal{F})$, $J^*_{[]}(\delta, \mathcal{F})$, and $\tilde{J}(\delta, \mathcal{F}, \|\cdot\|)$ of Section 11.1.

LEMMA 21.8 *(Entropy control with covering number) Let $\mathcal{F}$ be a $P$-measurable class of measurable functions with measurable envelope $F$. Let $\xi$ be a positive random variable with $\mathrm{E}(\xi) = 1$ and $\mathrm{var}(\xi) = v_0$, for $0 \leq v_0 < \infty$. Then*

$$\mathrm{E}^* \left[ \|\mathbb{G}_n(\xi f)\|_{\mathcal{F}} \right] \leq k J^*(1, \mathcal{F}) \|F\|_{P,2},$$

*where $k$ does not depend on $\mathcal{F}$, $\xi$ or $F$, and where $\|\cdot\|_{\mathcal{F}}$ denotes the supremum over all $f \in \mathcal{F}$.*

**Proof.** Since $\|\xi(f_1 - f_2)\|_{Q,2} \leq \|\xi\|_{Q,2} \|f_1 - f_2\|_{Q,2}$, we have

$$N\left(\epsilon \|\xi\|_{Q,2} \|F\|_{Q,2}, \xi\mathcal{F}, L_2(Q)\right) \leq N\left(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)\right). \tag{21.16}$$

Also, since $F$ is measurable, we have that $\xi F$ is measurable.

Set $\mathbb{G}_n^{\cdot} = n^{-1/2}\sum_{i=1}^n \epsilon_i f(X_i)$, where the $\epsilon_i$'s are i.i.d. Rademacher random variables, defined by $\mathrm{P}(\epsilon = 1) = \mathrm{P}(\epsilon = -1) = 1/2$. Set $\psi_2(y) = e^{y^2} - 1$. Then by Theorem 8.4, we have

$$\| \|\mathbb{G}_n^{\cdot}\|_{\xi\mathcal{F}}\|_{\psi_2|X,\xi} \lesssim \int_0^{\tau_n} \sqrt{1 + \log N(\epsilon, \xi\mathcal{F}, L_2(\mathbb{P}_n))}d\epsilon, \tag{21.17}$$

for $\tau_n = \| \|\xi f\|_n\|_{\mathcal{F}}$, where "$\lesssim$" means "bounded above up to a universal constant". Here $\|\cdot\|_n$ is the $L_2(\mathbb{P}_n)$-seminorm and $\|\cdot\|_{\psi_2}$ is the usual $\psi_2$ Orlicz norm. Setting $\epsilon = u\|\xi\|_n\|F\|_n$ and changing variables in (21.17), we obtain

$$\begin{aligned}
&\int_0^{\tau_n} \sqrt{1 + \log N(\epsilon, \xi\mathcal{F}, L_2(\mathbb{P}_n))}d\epsilon \\
&= K\int_0^{\frac{\tau_n}{\|\xi\|_n\|F\|_n}} \sqrt{1 + \log N(u\|\xi\|_n\|F\|_n, \xi\mathcal{F}, L_2(\mathbb{P}_n))}\|\xi\|_n\|F\|_n du \\
&\le K\int_0^1 \sqrt{1 + \log N(u\|F\|_n, \mathcal{F}, L_2(\mathbb{P}_n))}\|\xi\|_n\|F\|_n du.
\end{aligned}$$

Hence we can conclude $E\|\mathbb{G}_n^o\|_{\xi\mathcal{F}} \lesssim J(1,\mathcal{F})(E\|F\|^2)^{1/2}$ after taking expectations. Thus the desired result holds by symmetrization (Theorem 8.8) and the fact that $\|\cdot\|_{\psi_2}$ dominates the associated $L_2$ norm.□

Lemma 21.9 *(Entropy control with bracketing I) Let $\mathcal{F}$ be a class of measurable functions with envelop $F$, let $\xi$ be as in Lemma 21.8, then*

$$\mathrm{E}^*\left[\|\mathbb{G}_n(\xi f)\|_{\mathcal{F}}\right] \le k\sqrt{1+v_0}J_{[]}^*(1,\mathcal{F})\|F\|_{P,2},$$

*where $k$ does not depend on $\mathcal{F}$, $\xi$ or $F$.*

Proof. By Theorem 11.2, we have

$$\mathrm{E}\left[\|\mathbb{G}_n(\xi f)\|_{\mathcal{F}}^*\right] \lesssim J_{[]}^*\left(1,\xi\mathcal{F}\right)\|\xi F\|_{P,2} = J_{[]}^*\left(1,\xi\mathcal{F}\right)\sqrt{1+v_0}\|F\|_{P,2}.$$

Since $\|\xi f_1 - \xi f_2\|_{P,2} = \sqrt{1+v_0}\|f_1 - f_2\|_{P,2}$ and $\|\xi F\|_{P,2} = \sqrt{1+v_0}\|F\|_{P,2}$, we have

$$N_{[]}\left(\epsilon\sqrt{1+v_0}\|F\|_{P,2}, \xi\mathcal{F}, L_2(P)\right) = N_{[]}\left(\epsilon\|F\|_{P,2}, \mathcal{F}, L_2(P)\right).$$

Thus $J_{[]}^*(1,\xi\mathcal{F}) = J_{[]}^*(1,\mathcal{F})$, and the desired result follows.□

Lemma 21.10 *(Entropy control with bracketing II) Let $\mathcal{F}$ be a class of measurable functions with $Pf^2 < \delta^2$ and $\|f\|_\infty \le M < \infty$. Also let $\xi$ be as defined in Lemma 21.8, except that we require $\xi \le c$ for some fixed constant $c < \infty$. Then*

$$\mathrm{E}^*\left[\|\mathbb{G}_n(\xi f)\|_{\mathcal{F}}\right] \le k\sqrt{1+v_0}\tilde{J}_{[]}(\delta,\mathcal{F},L_2(P))\left\{1 + \frac{\tilde{J}_{[]}(\delta,\mathcal{F},L_2(P))}{\sqrt{1+v_0}\delta^2\sqrt{n}}cM\right\},$$

*where $k$ does not depend on $\mathcal{F}$, $\xi$ or $F$.*

PROOF. By Theorem 11.3, we have

$$\mathrm{E}^*\|\mathbb{G}_n\|_{\xi\mathcal{F}} \lesssim \tilde{J}_{[]}(\delta\sqrt{1+v_0}, \xi\mathcal{F}, L_2(P)) \left(1 + \frac{\tilde{J}_{[]}(\delta\sqrt{1+v_0}, \xi\mathcal{F}, L_2(P))}{(1+v_0)\delta^2\sqrt{n}} cM\right).$$

We also have $\|\xi\mathcal{F}\|_{P,2} \leq \sqrt{1+v_0}\delta$, $\|\xi\mathcal{F}\|_\infty \leq cM$, and $\|\xi f_1 - \xi f_2\|_{P,2} = \sqrt{1+v_0}\|f_1 - f_2\|_{P,2}$. From the definition of $\tilde{J}_{[]}$, we have

$$\begin{aligned}
\tilde{J}_{[]}(\delta, \xi\mathcal{F}, L_2(P)) &= \int_0^\delta \sqrt{1 + \log N_{[]}(\epsilon, \xi\mathcal{F}, L_2(P))}d\epsilon \\
&= \int_0^\delta \sqrt{1 + \log N_{[]}\left(\frac{\epsilon}{\sqrt{1+v_0}}, \mathcal{F}, L_2(P)\right)}d\epsilon \\
&= \int_0^{\frac{\delta}{\sqrt{1+v_0}}} \sqrt{1 + \log N_{[]}(u, \mathcal{F}, L_2(P))}\sqrt{1+v_0}du \\
&= \sqrt{1+v_0}\tilde{J}_{[]}\left(\frac{\delta}{\sqrt{1+v_0}}, \mathcal{F}, L_2(P)\right).
\end{aligned}$$

Hence the desired result follows.□

We note that these maximal inequalities all permit $v_0 = 0$. Hence the setting where $\xi = 1$ almost surely is a special case of the above lemmas. More precisely, the conclusions of Theorems 11.1, 11.2 and 11.3 are respectively implied by the above lemmas.

The previous three lemmas are enough for our examples, but there are many settings where other maximal inequalities may be needed. For a more complete reference on maximal inequalities without the random weights, see Section 11.1 of this book, Chapter 6 of van de Geer (2000), and Sections 2.14 and 3.4 of VW.

## 21.4 Examples Continued

We now study the models presented in Section 21.1.1 in greater detail. We only include relevant model assumptions here. For demonstration purposes and for clarity, some assumptions are made stronger here than in the original references.

### *21.4.1 Cox Model with Current Status Data (Example 1, Continued)*

The MLE of the Cox model with current status data was studied extensively in Section 19.2.3, and elsewhere in Chapter 19, and in both Huang (1996) and Section 25.11.1 of van der Vaart (1998). The OLS has not been studied previously. The main assumptions include:

E1.1: $\theta \in \Theta$ and $Z \in B$, where $\Theta$ and $B$ are known compact subsets of $\mathbb{R}^k$.

E1.2: There exists a known $K$, such that $0 < 1/K < \Lambda < K < \infty$.

E1.3: The event time $T$ and censoring time $Y$ are both bounded by a known constant, with $Y \in [\sigma, \tau]$, where $0 < \sigma < \tau < \infty$.

E1.4: The event time and censoring time are conditionally independent given $Z$.

We need the following technical tool. For a proof (which we omit), see Lemma 25.84 of van der Vaart (1998):

LEMMA 21.11 *Under Conditions E1.1–E1.4, there exists a constant $C$ such that, for every $\epsilon > 0$, and for*

$$\begin{aligned} \mathbb{M}_1 &= \{\delta \log\left(1 - \exp[-\exp(\theta' Z)\Lambda]\right) \\ &\quad -(1-\delta)\exp(\theta' Z)\Lambda : \theta \in \Theta,\, \Lambda \in \mathcal{L}\}, \text{ and} \\ \mathbb{M}_2 &= \left\{\left(1 - \delta - \exp[-\exp(\theta' Z)\Lambda]\right)^2 : \theta \in \Theta,\, \Lambda \in \mathcal{L}\right\}, \end{aligned}$$

*where $\mathcal{L}$ denotes all nondecreasing functions $\Lambda$ with $1/K \leq \Lambda(\sigma) \leq \Lambda(\tau) \leq K$, we have*

$$\log N_{[]}\left(\epsilon, \mathbb{M}_j, L_2(P)\right) \leq C\left(\frac{1}{\epsilon}\right), \text{ for } j = 1, 2. \tag{21.18}$$

*Consistency.*

The parameter set for $\theta$ is compact by assumption, and the parameter set for $\Lambda$ is compact relative to the weak topology (the topology used for weak convergence of measures, as done in Part (viii) of Theorem 7.6). Consistency for the MLE and OLS can be obtained by the argmax theorem (Theorem 14.1), as discussed for the MLE in Section 25.11.1 of van der Vaart (1998). The consistency for the OLS is quite similar, and we omit the details.

*Rate of convergence.*

Define $d\left((\theta, \Lambda), (\theta_0, \Lambda_0)\right) = |\theta - \theta_0| + \|\Lambda - \Lambda_0\|_2$, where $\|\cdot\|_2$ is the $L_2$ norm on $[\sigma, \tau]$. Combining the Inequality (21.18) from Lemma 21.11 with Lemma 21.9 above and taking $\phi_n(\delta) = \sqrt{\delta}\left(1 + \sqrt{\delta}/(\delta^2\sqrt{n})\right)$ in Corollary 21.4 above, we can conclude a convergence rate of $n^{1/3}$ for both $|\hat{\theta}_n - \theta_0|$ and $\|\hat{\Lambda}_n - \Lambda_0\|_2$. This result holds for both the MLE and OLS, since they have the same consistency and entropy results. The $n^{1/3}$ convergence rate is optimal for the estimation of $\Lambda$, as discussed in Groeneboom and Wellner (1992).

*$\sqrt{n}$ consistency and asymptotic normality.*

For the MLE, define $\tilde{m}(\theta, \Lambda) = \left\{ Z\Lambda - \phi(\Lambda)(h_{00} \circ \Lambda_0^{-1})\Lambda \right\} Q(X; \theta, \Lambda)$, where $Q$, $\phi$ and $h_{00}$ are as defined in Section 19.2.3. The following "finite variance" assumption is needed:

E1.5: $P[(Z\Lambda_0(Y) - h_{00}(Y))Q(X; \theta_0, \Lambda_0)]^{\otimes 2}$ is non-singular.

$\sqrt{n}$ consistency and asymptotic normality of $\hat{\theta}_n$ can then be proved by the "efficient score" approach given in Section 25.11.1 of van der Vaart. This is a special case of Theorem 21.1.

For the OLS, on the other hand, set

$$m_1 = 2Ze^{\theta' Z}\Lambda(Y)\exp(-e^{\theta' Z}\Lambda(Y))(1 - \delta - \exp(-e^{\theta' Z}\Lambda(Y))), \text{ and}$$
$$m_2[a] = 2e^{\theta' Z}\exp(-e^{\theta' Z}\Lambda(Y))(1 - \delta - \exp(-e^{\theta' Z}\Lambda(Y)))a(Y),$$

where $a \in \mathbb{A}$, and $\mathbb{A} = L_2(Y)$, the set of all mean zero, square-integrable functions of $Y$. Consider the estimator $(\hat{\theta}_n, \hat{\eta}_n)$ satisfying

$$\mathbb{P}_n m_1(\hat{\theta}_n, \hat{\eta}_n) = o_P(n^{-1/2}) \text{ and } \mathbb{P}_n m_2(\hat{\theta}_n, \hat{\eta}_n)[a] = o_P(n^{-1/2}),$$

for all $a \in \mathbb{A}$. We now prove the $\sqrt{n}$ consistency and asymptotic normality of $\hat{\theta}_n$ using Corollary 21.2. From the rate of convergence results discussed above, Condition A1 is satisfied with $c_1 = 1/3$. Condition B3 can be verified with the entropy result (21.18) and Lemma 21.10. Combining Taylor expansion with the differentiability of the functions involved in $m_1$ and $m_2$, we can see that Condition B4 is satisfied with $c_2 = 2$. Now we check Condition A2.

Define $\Lambda_t = \Lambda_0 + ta$ for a proper perturbation direction $a$. Set

$$\begin{aligned} m_{12}[a] &= 2Ze^{\theta' Z}\exp(-e^{\theta' Z}\Lambda) \\ &\quad \times \left( (1 - \Lambda e^{\theta' Z})(1 - \delta - \exp(-e^{\theta' Z}\Lambda)) + \Lambda e^{\theta' Z}\exp(-e^{\theta' Z}\Lambda) \right) a \\ &\equiv L(\theta, \Lambda)a \end{aligned}$$

and

$$\begin{aligned} m_{22}[a_1][a_2] &= -2e^{2\theta' Z}\exp(-e^{\theta' Z}\Lambda)\left[1 - \delta - 2\exp(-e^{\theta' Z}\Lambda)\right] a_1 a_2 \\ &\equiv R(\theta, \Lambda)a_1 a_2, \end{aligned}$$

for $a_1, a_2 \in \mathbb{A}$. Condition A2 requires $P(m_{12}[A] - m_{22}[A^*][A]) = 0$, for all $A \in \mathbb{A}^k$, which will be satisfied by $A^* \equiv \mathrm{E}(L(\theta, \Lambda)|Y)/\mathrm{E}(R(\theta, \Lambda)|Y)$. Simple calculations give

$$\begin{aligned} m_{11} &= 2Z^{\otimes 2}\Lambda e^{\theta' Z}\exp(-e^{\theta' Z}\Lambda) \\ &\quad \times \left[ (1 - \Lambda e^{\theta' Z})(1 - \delta - \exp(-e^{\theta' Z}\Lambda)) + \Lambda e^{\theta' Z}\exp(-e^{\theta' Z}\Lambda) \right] \end{aligned}$$

and $m_{21}[a] = L(\theta, \Lambda)a$. Define

$$I^* = \{P(m_{11} - m_{21}[A^*])\}^{-1} P[m_1 - m_2[A^*]]^{\otimes 2} \{P(m_{11} - m_{21}[A^*])\}^{-1},$$

and assume

E1.5′: $0 < \det(I^*) < \infty$.

Then all conditions of Corollary 21.2 are satisfied, and the desired $\sqrt{n}$ consistency and asymptotic normality of $\hat{\theta}_n$ follows.

*Validity of the weighted bootstrap.*

For the random weights $\xi$, we assume

E1.6: There exists a constant $c$ such that $\xi < c < \infty$.

Consistency of the weighted M-estimators can be established by Corollary 21.3, following the same arguments as for the ordinary M-estimators. With Condition E1.6 and Lemma 21.10, we can apply Corollary 21.4 to establish a convergence rate of $n^{1/3}$ for all parameters.

For the $\sqrt{n}$ consistency and asymptotic normality of $\hat{\theta}_n^\circ$, Corollaries 21.5 and 21.6 are applied to the weighted MLE and the weighted least squares estimator, respectively. For both estimators, Conditions A2 and A4 have been verified previously. We now only need to check Conditions A1 and A3 for the MLE and A1 and B3 for the least squares estimator. Condition A1 is satisfied with $c_1 = 1/3$ from the convergence rate results for both estimators. Conditions A3 and B3 can be checked directly with the entropy result (21.18) and Lemma 21.10. Hence the $\sqrt{n}$ consistency and asymptotic normality of the weighted estimators for $\theta$ follow. Based on Theorem 21.7, the validity of the weighted bootstrap follows for both the MLE and OLS.

### *21.4.2 Binary Regression Under Misspecified Link Function (Example 2, Continued)*

Denote the true value of $(\theta,\ h)$ as $(\theta_0,\ h_0)$. Under misspecification, it is assumed that

$$P_{\theta,h}(Y = 1 | Z = z, U = u) = \psi(\theta' z + h(u)), \tag{21.19}$$

where $\psi \neq \phi$ is an incorrect link function. When the model is misspecified, the maximizer $(\tilde{\theta}, \tilde{h})$ of the likelihood function is not necessarily $(\theta_0, h_0)$. The maximizer $(\tilde{\theta}, \tilde{h})$ can also be viewed as a minimizer of the Kullback-Leibler discrepancy, which is defined as

$$-P\left[\frac{\log P_{\theta,h}}{\log P_0}\right] = -P\left[\frac{Y \log \psi(\theta' Z + h) + (1 - Y) \log(1 - \psi(\theta' Z + h))}{Y \log \phi(\theta_0' Z + h_0) + (1 - Y) \log(1 - \phi(\theta_0' Z + h_0))}\right].$$

Here the expectation is taken with respect to the true underlying distribution. Thus $(\tilde{\theta}, \tilde{h})$ can be viewed as a parameter of the true distribution $P$.

The following model assumptions are needed.

E2.1: $\left(\frac{\psi^{(1)}}{\psi}\right)^{(1)} < 0$, $\left(-\frac{\psi^{(1)}}{1-\psi}\right)^{(1)} < 0$ and $\left(\frac{\phi^{(1)}}{\phi}\right)^{(1)} < 0$, $\left(-\frac{\phi^{(1)}}{1-\phi}\right)^{(1)} < 0$.

E2.2: For simplicity, we assume $U \in [0,1]$, and $E\tilde{h}(U) = c$ for a known constant $c$. We estimate $h$ under the constraint $\mathbb{P}_n(\hat{h}_n) = c$.

E2.3: $\theta \in B_1$ and $Z \in B_2$, where $B_1$ and $B_2$ are both compact sets in $\mathbb{R}^k$.

E2.4: $\mathrm{var}(Z - \mathrm{E}(Z|U))$ is non-singular.

The Condition E2.1 is satisfied by many link functions, including the logit link (corresponding to $\psi(u) = e^u/(1+e^u)$), the probit link (corresponding to $\psi(u) = \Phi(u)$), and the complementary log-log link (corresponding to $\psi(u) = \exp(-e^u)$).

Combining the second entropy result from Theorem 9.21 with the conditions that $\theta$ and $Z$ are both bounded and $\psi$ is continuously differentiable, we can conclude

$$\log N(\epsilon, \{\psi(\theta' Z + h(U))\}, L_2(Q)) \leq C\epsilon^{-1/s}, \tag{21.20}$$

where $h \in \mathbb{H}$, $\theta \in B_1$, $Z \in B_2$, all probability measures $Q$, and where $C$ is a fixed constant. Equation (21.20), together with the boundedness conditions and the Lipschitz property of the log function, leads to a similar result for the class of log-likelihood functions.

*Consistency.*

Let $m$ denote the log-likelihood. Combining $\mathbb{P}_n m(\hat{\theta}_n, \hat{h}_n) \geq \mathbb{P}_n m(\tilde{\theta}, \tilde{h})$ with $Pm(\hat{\theta}_n, \hat{h}_n) \leq Pm(\tilde{\theta}, \tilde{h})$, we have

$$0 \leq P(m(\tilde{\theta}, \tilde{h}) - m(\hat{\theta}_n, \hat{h}_n)) \leq (\mathbb{P}_n - P)\{m(\hat{\theta}_n, \hat{h}_n) - m(\tilde{\theta}, \tilde{h})\}.$$

The entropy result (21.20), boundedness Assumptions E2.2 and E2.3 and Lemma 21.8 give $\sqrt{n}(\mathbb{P}_n - P)\{m(\hat{\theta}_n, \hat{h}_n) - m(\tilde{\theta}, \tilde{h})\} = o_P(1)$. Thus we can conclude

$$0 \leq P(m(\tilde{\theta}, \tilde{h}) - m(\hat{\theta}_n, \hat{h}_n)) \leq o_P(n^{-1/2}).$$

Also, via Taylor expansion,

$$\begin{aligned} & P(m(\tilde{\theta}, \tilde{h}) - m(\hat{\theta}_n, \hat{h}_n)) \\ &= -P\left\{ \left((\tilde{\theta}' Z + \tilde{h}) - (\hat{\theta}_n' Z + \hat{h}_n)\right)^2 \left(\frac{\psi^{(1)}(\bar{\theta}' Z + \bar{h})}{\psi(\bar{\theta}' Z + \bar{h})}\right)^{(1)} Y \right. \\ &\quad \left. - \left(\frac{\psi^{(1)}(\bar{\theta}' Z + \bar{h})}{1 - \psi(\bar{\theta}' Z + \bar{h})}\right)^{(1)} (1 - Y) \right\}, \end{aligned}$$

where $(\bar{\theta}, \bar{h})$ is on the line segment between $(\hat{\theta}_n, \hat{h}_n)$ and $(\tilde{\theta}, \tilde{h})$.

Combining Condition E2.1 with Conditions E2.2 and E2.3, we know there exists a scalar $c_0$ such that

$$-\left\{\left(\frac{\psi^{(1)}(\bar{\theta}'Z+\bar{h})}{\psi(\bar{\theta}'Z+\bar{h})}\right)^{(1)} Y - \left(\frac{\psi^{(1)}(\bar{\theta}'Z+\bar{h})}{1-\psi(\bar{\theta}'Z+\bar{h})}\right)^{(1)} (1-Y)\right\} > c_0 > 0.$$

Hence we can conclude $P\left((\tilde{\theta}'Z+\tilde{h}) - (\hat{\theta}_n'Z+\hat{h}_n)\right)^2 = o_P(1)$, which is equivalent to

$$P\left\{(\tilde{\theta}-\hat{\theta}_n)'(Z-\mathrm{E}(Z|U)) + (\tilde{h}-\hat{h}_n) + (\tilde{\theta}-\hat{\theta}_n)'\mathrm{E}(Z|U)\right\}^2 = o_P(1).$$

Since $\mathrm{var}(Z-\mathrm{E}(Z|U))$ is non-singular, the above equation gives us $|\hat{\theta}_n - \tilde{\theta}| = o_P(1)$ and $\|\hat{h}_n - \tilde{h}\|_2 = o_P(1)$.

*Rate of convergence.*

Define $d((\theta_1, h_1), (\theta_2, h_2)) = |\theta_1 - \theta_2| + \|h_1 - h_2\|_{L_2}$. From the boundedness assumptions, the log-likelihood functions are uniformly bounded. Based on the entropy result given in Theorem 9.21 and Lemma 21.8, and using the bound of the log-likelihood function as the envelope function, we now have

$$\mathrm{E}^*|\sqrt{n}(\mathbb{P}_n - P)(m(\tilde{\theta}, \tilde{h}) - m(\hat{\theta}_n, \hat{h}_n))| \leq K_1 \delta^{1-1/(2s)}, \tag{21.21}$$

for $d((\tilde{\theta}, \tilde{h}), (\hat{\theta}_n, \hat{h}_n)) < \delta$ and a constant $K_1$. This result can then be applied to Corollary 21.4 above to yield the rate of convergence $n^{s/(2s+1)}$.

*$\sqrt{n}$ consistency and asymptotic normality.*

Corollary 21.2 is applied here. We can see that Condition A1 is satisfied with $c_1 = s/(2s+1)$; Condition B3 has been shown in (21.21); and Condition B4 is satisfied with $c_2 = 2$ based on the continuity of the involved functions and on Taylor expansion. Now we only need to verify the finite variance condition.

We have

$$\begin{aligned} m_1 &= \left(Y\frac{\psi^{(1)}}{\psi} - (1-Y)\frac{\psi^{(1)}}{1-\psi}\right) Z \text{ and} \\ m_2[a] &= \left(Y\frac{\psi^{(1)}}{\psi} - (1-Y)\frac{\psi^{(1)}}{1-\psi}\right) a, \end{aligned}$$

for a proper tangent $a$, and where $h_t = h_0 + ta$. We also have for $a_1, a_2 \in \mathbb{A}$, where $\mathbb{A} = \{a : a \in L_2^0(U) \text{ with } J(a) < \infty\}$, that

$$
\begin{aligned}
m_{11} &= \left( Y \left( \frac{\psi^{(1)}}{\psi} \right)^{(1)} - (1-Y) \left( \frac{\psi^{(1)}}{1-\psi} \right)^{(1)} \right) Z^{\otimes 2}, \\
m_{12}[a_1] &= \left( Y \left( \frac{\psi^{(1)}}{\psi} \right)^{(1)} - (1-Y) \left( \frac{\psi^{(1)}}{1-\psi} \right)^{(1)} \right) Z a_1, \\
m_{21}[a_1] &= \left( Y \left( \frac{\psi^{(1)}}{\psi} \right)^{(1)} - (1-Y) \left( \frac{\psi^{(1)}}{1-\psi} \right)^{(1)} \right) Z a_1, \quad \text{and} \\
m_{22}[a_1][a_2] &= \left( Y \left( \frac{\psi^{(1)}}{\psi} \right)^{(1)} - (1-Y) \left( \frac{\psi^{(1)}}{1-\psi} \right)^{(1)} \right) a_1 a_2.
\end{aligned}
$$

The "finite variance" condition requires that there exists $A^* = (a_1^* \dots a_k^*)$, such that for any $A = (a_1 \dots a_k)$

$$
\begin{aligned}
(21.22)\ & P\{m_{12}[A] - m_{22}[A^*][A]\} \\
&= P\left\{ \left( Y \left( \frac{\psi^{(1)}}{\psi} \right)^{(1)} - (1-Y) \left( \frac{\psi^{(1)}}{1-\psi} \right)^{(1)} \right) (ZA' - A^*A') \right\} \\
&= 0,
\end{aligned}
$$

where $a_i^*, a_i \in \mathbb{A}$. Denote $Q \equiv \left( Y \left( \frac{\psi^{(1)}}{\psi} \right)^{(1)} - (1-Y) \left( \frac{\psi^{(1)}}{1-\psi} \right)^{(1)} \right)$. It is clear that $A^* = E(ZQ|U)/E(Q|U)$ satisfies (21.22). Assume

E2.5:

$$
\begin{aligned}
0 &< \det\left( \{P(m_{11} - m_{21}[A^*])\}^{-1} P[m_1 - m_2[A^*]]^{\otimes 2} \right. \\
&\quad \left. \times \{P(m_{11} - m_{21}[A^*])\}^{-1} \right) \\
&< \infty.
\end{aligned}
$$

Now the $\sqrt{n}$ consistency and asymptotic normality of $\hat{\theta}_n$ follow.

*Validity of the weighted bootstrap.*

Consistency of the weighted MLE can be obtained similarly to that for the ordinary MLE, with reapplication of the entropy result from Lemma 21.8. The rate of convergence can be obtained by Corollary 21.4 as $n^{s/(2s+1)}$ for all parameters. Unconditional $\sqrt{n}$ consistency and asymptotic normality for the estimator of $\tilde{\theta}$ can be achieved via Corollary 21.6. Thus from Theorem 21.7, the weighted bootstrap is valid.

### *21.4.3 Mixture Models (Example 3, Continued)*

Consider the mixture model with kernel $p_\theta(X, Y|Z) = Ze^{-ZX}\theta Ze^{-\theta ZY}$, where we write the observations as pairs $(X_i, Y_i)$, $i = 1, \dots, n$. Thus given

unobservable variables $Z_i = z$, each observation consists of a pair of exponentially distributed variables with parameters $z$ and $\theta z$, respectively. Assume $Z$ has unknown distribution $\eta(z)$.

Existence of the "least-favorable direction" $a^*$ for this model is discussed in Section 3 of van der Vaart (1996). The efficient score function is shown to be

$$\tilde{\ell}_{\theta,\eta}(x,y) = \frac{\int \frac{1}{2}(x-\theta y)z^3 \exp(-z(x+\theta y))d\eta(z)}{\int \theta z^2 \exp(-z(x+\theta y))d\eta(z)}. \tag{21.23}$$

We assume the following conditions hold:

E3.1: $\theta \in B$, where $B$ is a compact set in $\mathbb{R}$.

E3.2: The true mixing distribution $\eta_0$ satisfies $\int (z^2 + z^{-5})d\eta_0(z) < \infty$.

E3.3: $P\{\tilde{\ell}_{\theta_0,\eta_0}\}^2 > 0$.

We need the following technical tool. The result (the proof of which we omit) comes form Lemma L.23 of Pfanzagl (1990).

LEMMA 21.12 *There exists a weak neighborhood $V$ of the true mixing distribution such that the class $\mathcal{F} \equiv \{\tilde{\ell}_{\theta,\eta} : \theta \in B, \eta \in V\}$ is Donsker with*

$$\log N_{[]}\left(\epsilon, \{\tilde{\ell}_{\theta,\eta} : \theta \in B, \eta \in V\}, L_2(P)\right) \leq K\left(\frac{1}{\epsilon}\right)^L$$

*for fixed constants $K < \infty$ and $1 < L < 2$.*

*Consistency.*

Consistency of $(\hat{\theta}_n, \hat{\eta}_n)$ for the product of the Euclidean and weak topology follows from Kiefer and Wolfowitz (1956), as discussed in van der Vaart (1996).

*Rate of convergence.*

From Lemma 21.12, we obtain that the entropy integral $\tilde{J}_{[]}(\epsilon, \mathcal{F}, L_2(P)) \lesssim \epsilon^{1-L/2}$. Lemma 21.10 and Corollary 21.4 can now be applied to obtain the convergence rate $n^{1/(2+L)}$ for all parameters.

*$\sqrt{n}$ consistency and asymptotic normality of $\hat{\theta}_n$.*

We now check the conditions of Theorem 21.1. Condition A1 is satisfied with $c_1 = 1/(2+L)$, as shown above. Condition A2 is the Assumption E3.3. See van der Vaart (1996) for a discussion of the tangent set $\mathbb{A}$. Condition A3 can be checked by Lemma 21.10 above. Condition A4 is satisfied with $c_2 = 2$. Hence the $\sqrt{n}$ consistency and asymptotic normality of $\hat{\theta}_n$ follow.

*Validity of the weighted bootstrap.*

We now establish properties of the weighted MLE. Consistency can be obtained following the arguments in Kiefer and Wolfowitz (1956), using the same arguments used for the ordinary MLE. The convergence rate $n^{1/(2+L)}$ can be achieved with Corollary 21.4 and the entropy result from Lemma 21.9 for all parameters. $\sqrt{n}$ consistency and asymptotic normality of the unconditional weighted MLE for $\theta$ can be obtained by Corollary 21.5. Thus we can conclude that the conditional weighted bootstrap is valid by Theorem 21.7.

## 21.5 Penalized M-estimation

Penalized M-estimators have been studied extensively, for example in Wahba (1990), Mammen and van de Geer (1997), van de Geer (2000, 2001), and Gu (2002). Generally, these estimators do not fit in the framework discussed above because of the extra penalty terms. Another important feature of penalized M-estimators is the difficulty of the inference. Although Murphy and van der Vaart (2000) show that inference for the semiparametric MLE can be based on the profile likelihood, their technique is not directly applicable to the penalized MLE. The main difficulty is that the penalty term of the objective function may have too large an order, and thus Condition (3.9) in Murphy and van der Vaart (2000) may not be satisfied in the limit.

We now show that certain penalized M-estimators can be coaxed into the framework discussed above. Examples we investigate include the penalized MLE for binary regression under misspecified link function, the penalized least squares estimator for partly linear regression, and the penalized MLE for the partly additive transformation model with current status data. The first example we will examine in detail, while the latter two examples we will only consider briefly. The key is that once we can establish that the penalty term is $o_P(n^{-1/2})$, then the "nearly-maximizing" condition of Corollary 21.4 (see also (21.12) and (21.6)) is satisfied. After this, all of the remaining analysis can be carried out in the same manner as done for ordinary semiparametric M-estimators.

### *21.5.1 Binary Regression Under Misspecified Link Function (Example 2, Continued)*

As discussed earlier, we can study this model under the relaxed condition $\int_0^1 \left(h^{(s)}(u)\right)^2 du < \infty$ by taking the penalized approach. All other assumptions as discussed in Section 21.4 for this model will still be needed. Consider the penalized MLE $(\hat{\theta}_n, \hat{h}_n)$ as an estimator of $(\tilde{\theta}, \tilde{h})$, where

$$(\hat{\theta}_n, \hat{h}_n) = \text{argmax}\{\mathbb{P}_n m(\theta, h) - \lambda_n^2 J^2(h)\}. \tag{21.24}$$

In (21.24), $\lambda_n$ is a data-driven smoothing parameter and

$$J^2(h) \equiv \int_0^1 \left(h^{(s)}(u)\right)^2 du$$

as defined previously. We also assume

E2.5: $\lambda_n = O_P(n^{-s/(2s+1)})$ and $\lambda_n^{-1} = O_P(n^{s/(2s+1)})$.

*Consistency.*

From the entropy result of Theorem 9.21 (with $m$ in the theorem replaced by $s$) we have

$$(\mathbb{P}_n - P)(m(\theta, h) - m(\tilde{\theta}, \tilde{h})) = (1 + J(h))O_P(n^{-1/2}), \tag{21.25}$$

where $\theta, h$ satisfy all the model assumptions (see Exercise 21.6.3). The penalized MLE satisfies

$$\mathbb{P}_n m(\hat{\theta}_n, \hat{h}_n) - \lambda_n^2 J^2(\hat{h}_n) \geq \mathbb{P}_n m(\tilde{\theta}, \tilde{h}) - \lambda_n^2 J^2(\tilde{h}). \tag{21.26}$$

Combining (21.25) and (21.26), we can conclude

$$\begin{aligned} 0 &\leq P(m(\tilde{\theta}, \tilde{h}) - m(\hat{\theta}_n, \hat{h}_n)) \\ &\leq \lambda_n^2 J^2(\tilde{h}) - \lambda_n^2 J^2(\hat{h}_n) + O_P(n^{-1/2})(1 + J(\hat{h}_n)). \end{aligned} \tag{21.27}$$

Under boundedness assumptions, we can now conclude $\lambda_n J(\hat{h}_n) = o_P(1)$. Hence $0 \leq P(m(\tilde{\theta}, \tilde{h}) - m(\hat{\theta}_n, \hat{h}_n)) = o_P(1)$. Following the same arguments as used in Section 21.4, we can now conclude the consistency of $(\hat{\theta}_n, \hat{h}_n)$.

*Rate of convergence.*

Denote $\psi = \theta' Z + h(U)$, $\tilde{\psi} = \tilde{\theta}' Z + \tilde{h}(U)$ and $\hat{\psi}_n = \hat{\theta}_n' Z + \hat{h}_n(U)$. Taylor expansion gives us

$$P(m(\tilde{\theta}, \tilde{h}) - m(\hat{\theta}_n, \hat{h}_n)) = P\left(\frac{1}{2} m^{(2)}(\bar{\psi})(\hat{\psi}_n - \tilde{\psi})^2\right), \tag{21.28}$$

where $\bar{\psi}$ is on the line segment between $\hat{\psi}_n$ and $\tilde{\psi}$. From the compactness assumptions, there exist $\epsilon_1$ and $\epsilon_2$, such that $0 < \epsilon_1 < m^{(2)} < \epsilon_2 < \infty$ almost surely. Combining this result with (21.28), we now have

$$\epsilon_1 \|\hat{\psi}_n - \tilde{\psi}\|^2 \leq P(m(\tilde{\theta}, \tilde{h}) - m(\hat{\theta}_n, \hat{h}_n)) \leq \epsilon_2 \|\hat{\psi}_n - \tilde{\psi}\|^2. \tag{21.29}$$

Careful application of Theorem 15.3, with $\alpha = 1/s$, combined with the fact that there exists a constant $k < \infty$ such that $|m(\hat{\psi}_n) - m(\tilde{\psi})| \leq k\|\hat{\psi}_n - \tilde{\psi}\|$ almost surely, yields

$$
\begin{aligned}
(21.30)\quad & |(\mathbb{P}_n - P)(m(\hat{\psi}_n) - m(\tilde{\psi}))| \\
&\leq O_P(n^{-1/2})(1 + J(\hat{h}_n)) \\
&\quad \times \left( \|\hat{\psi}_n - \tilde{\psi}\|^{1-1/(2s)} \vee n^{-(2s-1)/[2(2s+1)]} \right)
\end{aligned}
$$

(see Exercise 21.6.4). Combining this with inequalities (21.29) and (21.26) and the boundedness assumptions, we have

$$
\begin{aligned}
(21.31)\quad & \lambda_n^2 J^2(\hat{h}_n) + \epsilon_1 \|\hat{\psi}_n - \tilde{\psi}\|^2 \\
&\leq \lambda_n^2 J^2(\tilde{h}) + O_P(n^{-1/2})(1 + J(\hat{h}_n)) \\
&\quad \times (\|\hat{\psi}_n - \tilde{\psi}\|^{1-1/(2s)} \vee n^{-(2s-1)/[2(2s+1)]}).
\end{aligned}
$$

Some simple calculations, similar to those used in Section 15.1, yield $J(\hat{h}_n) = O_P(1)$ and $\|\hat{\psi}_n - \tilde{\psi}\| = O_P(n^{-s/(2s+1)})$ (see Exercise 21.6.5). Once these results are established, the remaining analysis for the ordinary MLE can be carried out as done in Section 21.4. We can see that, if we replace $\mathbb{P}_n$ with $\mathbb{P}_n^\circ$ in the above analysis, all results hold with only minor modifications needed. Thus we can establish the consistency and rate of convergence results for the weighted MLE similarly. Then the analysis of the weighted MLE for $\tilde{\theta}$ and the validity of the weighted bootstrap can be carried out using the same arguments as used in Section 21.4.

### *21.5.2 Two Other Examples*

Two other penalized M-estimation examples are studied in in detail Ma and Kosorok (2005b). The first example is partly linear regression, where the observed data consist of $(Y, Z, U)$, where $Y = \theta' Z + h(U) + \epsilon$, where $h$ is nonparametric with $J(h) < \infty$ and $\epsilon$ is normally distributed. Penalized least-squares is used for estimation and the weighted bootstrap is used for inference. The main result is that the penalized bootstrap $\hat{\theta}_n^*$ is valid, i.e., that $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) \underset{\xi}{\overset{\mathrm{P}}{\rightsquigarrow}} Z$, where $Z$ is the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$.

The second example is for partly linear transformation models for current status data. This model was also studied in Ma and Kosorok (2005a). The observed data is $(V, 1\{U \leq V\}, Z, W)$, where $U$ is the failure time of interest, $V$ is a random observation time, $Z$ is a Euclidean covariate and $W$ is a continuous covariate. The model is $H(U) + \theta' Z + h(W) = e$, where $e$ has a known distribution, and $U$ and $V$ are assumed to be independent given $(Z, W)$. The nuisance parameter $\eta = (H, h)$ consists of two components, a transformation function $H$ and a nonparametric regression function $h$. It is assumed that $H$ is monotone and $J(h) < \infty$. Only $h$ is penalized in the maximum likelihood. The main result is, as in the previous paragraph, that the conditional weighted bootstrap is valid for inference on $\theta$. Additional details can be found in Ma and Kosorok (2005b).

## 21.6 Exercises

**21.6.1.** Prove Corollaries 21.3 and 21.4. This can be done by clarifying how the presence of random weights affects the arguments used in the proofs of Theorems 2.12 and 14.4. Lemma 14.3 may also be helpful for proving Corollary 21.3.

**21.6.2.** Prove Corollaries 21.5 and 21.6 using the arguments in the respective proofs of Theorem 21.1 and Corollary 21.2.

**21.6.3.** Verify that (21.25) holds. Hint: Consider the class

$$\mathcal{F} = \left\{ \frac{m(\theta, h) - m(\tilde{\theta}, \tilde{h})}{1 + J(h)} : \theta \in \Theta, J(h) < \infty \right\},$$

and verify that Theorem 9.21 can be applied, after some manipulations, to obtain

$$\sup_Q \log N(\epsilon, \mathcal{F}, L_2(Q)) \leq M_* \left(\frac{1}{\epsilon}\right)^{1/s},$$

where the supremum is over all finitely-discrete probability measures and $M_* < \infty$.

**21.6.4.** Verify that (21.30) holds.

**21.6.5.** In the context of Section 2.5.1, verify that $J(\hat{h}_n) = O_P(1)$ and $\|\hat{\psi}_n - \tilde{\psi}\| = O_P(n^{-2s/(2s+1)})$. Hint: Let $A_n \equiv n^{s/(2s+1)}\|\hat{\psi}_n - \tilde{\psi}\|$, and show that (21.31) implies

$$J^2(\hat{h}_n) + A_n^2 = O_P(1) + O_P(1)(1 + J(\hat{h}_n)) \left(A_n^{1-1/(2s)} \vee 1\right).$$

Thus $A_n^2 = O_P(1) + O_P(1)(1 + J(\hat{h}_n))A_n^{1-1/(2s)}$, and hence $J^2(\hat{h}_n) = O_P(1)(1+J(\hat{h}_n))^{4s/(2s+1)}$ (this step is a little tricky). Thus $J(\hat{h}_n) = O_P(1)$, and the desired result can now be deduced.

## 21.7 Notes

As mentioned previously, the material in this chapter is adapted from Ma and Kosorok (2005b). In particular, Examples 1–3 are the first three examples in Ma and Kosorok, and many of the theorems, corollaries and lemmas of this chapter also have corresponding results in Ma and Kosorok. Theorem 21.1 and Corollaries 21.2 through 21.5 correspond to Theorem 1, and Corollaries 1, 3, 4, and 2, respectively, of Ma and Kosorok. Corollary 21.6 is based on remark 7 of Ma and Kosorok; and Theorem 21.7 and Lemmas 21.8 through 21.12 correspond to Theorem 2, Lemmas 1–3, and Technical Tools T1 and T3, respectively, of Ma and Kosorok.

# 22
# Case Studies III

In this chapter, we examine closely four examples that utilize or are related to some of the concepts discussed in Part III. In most cases, concepts and results from Part II also play a key role. A secondary goal in these examples is to illustrate a paradigm for statistical research that can be applied to many different data generating models. This is particularly true for the first example on the proportional odds model under right censoring. This model has been discussed extensively in previous chapters, but it is helpful to organize all the steps that are (or should be) taken in developing inference for this model. A few holes in these steps are filled in along the way.

The second example is a follow-up on the linear regression model introduced in Chapter 1 and discussed in Section 4.1.1. The objective is to develop efficient estimation for the setting where the covariate and residual are known to be independent. The third example is new and is concerned with a time-varying regression model for temporal process data. An interesting issue for functional data—such as temporal process data—is that sometimes optimal efficiency is not possible, although improvements in efficiency within certain classes of estimating equations are possible. Empirical process methods play a crucial role in developing methods of inference for this setting.

The fourth example, also new, considers partly linear regression for repeated measures data. The dependencies between the repeated measures in a single observation produce nontrivial complications in the asymptotic theory. Other complications arise from the lack of boundedness restrictions on parameter spaces which is in contrast to the assumptions made for other

partly linear regression models considered in this book (see, for example, the first paragraph of Section 15.1).

These examples, in combination with other examples presented previously in this book, demonstrate clearly the power of semiparametric and empirical process methods to enlarge the field of possibilities in scientific research for both flexible and meaningful statistical modeling. It is hoped that the reader will have not only have gained a meaningful technical background that will facilitate further learning and research, but, more importantly, it is hoped that the reader's imagination has been stimulated with new ideas and insights.

## 22.1 The Proportional Odds Model Under Right Censoring Revisited

The proportional odds model for right censored data has been studied extensively in Sections 15.3 and 19.2.2 and several places in Chapter 20, including Exercise 20.3.4. The main steps for estimation and inference that have been discussed for this model are:

- Developing a method of estimation. A maximum log-likelihood approach was developed and discussed in Section 15.3.1.
- Establishing consistency of the estimator. This was accomplished in Sections 15.3.2 and 15.3.3. A preliminary step of establishing existence (Section 15.3.2) was also needed in this instance for the baseline function $A$ because of the omission of boundedness assumptions on $A$ (which omission was done to be most realistic in practice).
- Establishing the rates of convergence. This step was addressed indirectly in this instance through weak convergence since all of the parameters end up being regular.
- Obtaining weak convergence for all regular parameters. This was accomplished in Sections 15.3.4 and 15.3.5. A Z-estimator approach was utilized that involved both score and information operators.
- Establishing efficiency of all regular parameters. This was accomplished for all parameters in the paragraphs following the presentation of Corollary 20.1. This step may be considered optional if maximum likelihood-like inference is difficult to implement or is impractical for some reason. Nevertheless, it is useful and often fruitful to at least attempt to obtain efficiency in estimation.
- Obtaining convergence of non-regular parameters. This step was not needed in this instance. In general, this step is optional if scientific interest is restricted to only the regular parameters.

- Developing a method of inference. The weighted bootstrap was suggested and validated in Section 15.3.5 for this model. For computational simplicity, the profile sampler of Chapter 19 and the piggyback bootstrap of Chapter 20 were also proposed for this model (see Section 19.2.2 and toward the end of Section 20.2). The only part that remains to be done is to verify all of the conditions of Theorems 19.5 and 19.6. This will establish the validity of the profile sampler. The validity of the piggyback bootstrap has already been established towards the end of Section 20.2.
- Studying the properties of estimation and inference under model misspecification. This is an optional, but often useful, step. This was accomplished to some degree for a generalization of the proportional odds model, the class of univariate frailty regression models, in Kosorok, Lee and Fine (2004).

Note that in practice, a researcher may need to iterate between several of these steps before achieving all of the desired conclusions. For instance, it may take a few iterations to arrive at a computationally feasible and efficient estimator, or, it may take a few iterations to arrive at the optimal rate of convergence. In any case, the above list of steps can be considered a general paradigm for estimation and inference, although there are probably other useful paradigms that could serve the same purpose.

The focus of the remainder of this section will be on establishing the conditions of Theorems 19.5 and 19.6 as mentioned above, since these are the only missing pieces in establishing the validity of the profile sampler for the proportional odds model under right censoring. As mentioned above, all of the remaining steps, except for these, have been addressed previously.

Recall from Section 19.2.2 that

$$A_t(\beta, A) = \int_0^{(\cdot)} (1 + (\beta - t)' h_0(s))\, dA(s),$$

where $h_0(s) = [\sigma^{22}_{\theta_0}]^{-1}\sigma^{21}_{\theta_0}(\cdot)(s)$ is as defined in that section, satisfies Conditions 19.2 and 19.3. Hence

$$\begin{aligned}
\dot{\ell}(t, \beta, A) &= \left(\frac{\partial}{\partial t}\right) l(t, A_t(\beta, A)) \\
&= \left(\frac{\partial}{\partial t}\right) \Big\{ \delta(\log \Delta A_t(\beta, A)(U) + t'Z) \\
&\quad -(1+\delta)\log\left(1 + e^{t'Z} A_t(\beta, A)(U)\right)\Big\} \\
&= \int_0^\tau \left(Z - \frac{h_0(s)}{1 + (\beta - t)' h_0(s)}\right) \\
&\quad \times \left[dN(s) - (1+\delta)\frac{Y(s) e^{t'Z} dA_t(\beta, A)(s)}{1 + e^{t'Z} A_t(\beta, A)(U \wedge \tau)}\right]
\end{aligned} \tag{22.1}$$

and

$$\ddot{\ell}(t,\beta,A) = \left(\frac{\partial}{\partial t}\right)\dot{\ell}(t,\beta,A) \tag{22.2}$$

$$= -\int_0^\tau \frac{h_0^{\otimes 2}(s)}{(1+(\beta-t)'h_0(s))^2}\left[dN(s) - (1+\delta)\frac{Y(s)e^{t'Z}dA_t(\beta,A)(s)}{1+e^{t'Z}A_t(\beta,A)(U\wedge\tau)}\right]$$

$$-(1+\delta)\int_0^\tau \left[Z - \frac{h_0(s)}{1+(\beta-t)'h_0(s)}\right]^{\otimes 2}\frac{Y(s)e^{t'Z}dA_t(\beta,A)(s)}{1+e^{t'Z}A_t(\beta,A)(U\wedge\tau)}$$

$$+(1+\delta)\left\{\int_0^\tau \left[Z - \frac{h_0(s)}{1+(\beta-t)'h_0(s)}\right]\frac{Y(s)e^{t'Z}dA_t(\beta,A)(s)}{1+e^{t'Z}A_t(\beta,A)(U\wedge\tau)}\right\}^{\otimes 2}.$$

Hence it is easy to see that both $(t,\beta,A) \mapsto \dot{\ell}(t,\theta,A)(X)$ and $(t,\beta,A) \mapsto \ddot{\ell}(t,\theta,A)(X)$ are continuous for $P_0$-almost every $X$. Although it is tedious, it is not hard to verify that for some uniform neighborhood $V$ of $(\beta_0,\beta_0,A_0)$,

$$\mathcal{F}_1 \equiv \left\{\dot{\ell}(t,\beta,A) : (t,\beta,A) \in V\right\} \tag{22.3}$$

is $P_0$-Donsker and

$$\mathcal{F}_2 \equiv \left\{\ddot{\ell}(t,\beta,A) : (t,\beta,A) \in V\right\} \tag{22.4}$$

is $P_0$-Glivenko-Cantelli (see Exercise 22.6.1).

Now consider any sequence $\tilde{\beta}_n \xrightarrow{\mathrm{P}} \beta_0$, and let $\hat{A}_{\tilde{\beta}_n}$ be the profile maximizer at $\beta = \tilde{\beta}_n$, i.e., $\hat{A}_{\tilde{\beta}_n} = \mathrm{argmax}_A L_n(A,\tilde{\beta}_n)$, where $L_n$ is the log-likelihood as defined in Section 15.3.1. The arguments in Section 15.3.2 can be modified to verify that $\hat{A}_{\tilde{\beta}_n}(\tau)$ is asymptotically bounded in probability. Since, by definition, $L_n(\hat{\beta}_n,\hat{A}_n) \geq L_n(\tilde{\beta}_n,\hat{A}_{\tilde{\beta}_n}) \geq L_n(\tilde{\beta}_n,\tilde{A}_n)$, where $\tilde{A}_n$ is as defined in Section 15.3.3, we can argue along the lines used in Section 15.3.3 to obtain that $\hat{A}_{\tilde{\beta}_n}$ is uniformly consistent for $A_0$.

We now wish to strengthen this last result to

$$\|\hat{A}_{\tilde{\beta}_n} - A_0\|_{[0,\tau]} = O_P\left(n^{-1/2} + \|\tilde{\beta}_n - \beta_0\|\right), \tag{22.5}$$

for any sequence $\tilde{\beta}_n \xrightarrow{\mathrm{P}} \beta_0$. If (22.5) holds, then, as shown by Murphy and van der Vaart (2000) in the discussion following their Theorem 1,

$$P\dot{\ell}(\beta_0,\tilde{\beta}_n,\hat{A}_{\tilde{\beta}_n}) = o_P\left(n^{-1/2} + \|\tilde{\beta}_n - \beta_0\|\right),$$

and thus both Conditions (19.11) and (19.12) hold. Hence all of the conditions of Theorem 19.5 hold.

We now show (22.5). The basic idea of the proof is similar to arguments given in the proof of Theorem 3.4 in Lee (2000). Recall the definition of

$V^{\tau}_{n,2}$ as given in Expression (15.10) and the space $\mathcal{H}^2_\infty$ from Section 19.2.2; and define, for all $h \in \mathcal{H}^2_\infty$,

$$\tilde{D}_n(A)(h) \equiv V^{\tau}_{n,2}(\tilde{\beta}_n, A)(h), \quad D_n(A)(h) \equiv V^{\tau}_{n,2}(\beta_0, A)(h),$$

and

$$D_0(A)(h) \equiv PV^{\tau}_{n,2}(\beta_0, A)(h).$$

By definition of a maximizer, $\tilde{D}_n(\hat{A}_{\tilde{\beta}_n})(h) = 0$ and $D_0(A_0)(h) = 0$ for all $h \in \mathcal{H}^2_\infty$.

By Lemma 13.3,

$$\sqrt{n}(\tilde{D}_n - D_0)(\hat{A}_{\tilde{\beta}_n}) - \sqrt{n}(\tilde{D}_n - D_0)(A_0) = o_P(1),$$

uniformly in $\ell^\infty(\mathcal{H}^2_1)$, where $\mathcal{H}^2_1$ is the subset of $\mathcal{H}^2_\infty$ consisting of functions of total variation $\leq 1$. By the differentiability of the score function, we also have

$$\sqrt{n}(\tilde{D}_n(A_0) - D_n(A_0)) = O_P(\sqrt{n}\|\tilde{\beta}_n - \beta_0\|),$$

uniformly in $\ell^\infty(\mathcal{H}^2_1)$. Combining the previous two displays, we have

$$\begin{aligned}
\sqrt{n}(D_0(\hat{A}_{\tilde{\beta}_n}) - D_0(A_0)) &= -\sqrt{n}(\tilde{D}_n(\hat{A}_{\tilde{\beta}_n}) - D_0(\hat{A}_{\tilde{\beta}_n})) \\
&= -\sqrt{n}(\tilde{D}_n - D_0)(A_0) + o_P(1) \\
&= -\sqrt{n}(D_n - D_0)(A_0) \\
&\quad + O_P(1 + \sqrt{n}\|\tilde{\beta}_n - \beta_0\|) \\
&= O_P(1 + \sqrt{n}\|\tilde{\beta}_n - \beta_0\|).
\end{aligned}$$

Note that $D_0(A)(h) = \int_0^\tau (\sigma^{22}_{\beta_0,A_0} h) dA(s)$, where $\sigma^{22}$ is as defined in Section 15.3.4. Since $\sigma^{22}_{\beta_0,A_0}$ is continuously invertible as shown in Section 19.2.2, we have that there exists some $c > 0$ such that $\sqrt{n}(D_0(\hat{A}_{\tilde{\beta}_n}) - D_0(A_0)) \geq c\|\hat{A}_{\tilde{\beta}_n} - A_0\|_{\mathcal{H}^2_1}$. Thus (22.5) is satisfied.

The only thing remaining to do for this example is to verify (19.17), after replacing $\tilde{\theta}_n$ with $\tilde{\beta}_n$ and $\Theta$ with $B$, since this well then imply the validity of the conclusions of Theorem 19.6. For each $\beta \in B$, let

$$A_\beta \equiv \text{argmax}_A P\left[\frac{dP_{\beta,A}}{dP_{\beta_0,A_0}}\right],$$

where $P = P_{\beta_0,A_0}$ by definition, and define

$$\tilde{\Delta}_n(\beta) \equiv P\left[\frac{dP_{\beta,A_\beta}}{dP_{\tilde{\beta}_n,\hat{A}_n}}\right] \quad \text{and} \quad \Delta_0(\beta) \equiv P\left[\frac{dP_{\beta,A_\beta}}{dP_{\beta_0,A_0}}\right].$$

Theorem 2 of Kosorok, Lee and Fine (2004) is applicable here since the proportional odds model is a special case of the odds rate model with frailty variance parameter $\gamma = 1$. Hence

$$\sup_{\beta\in B} |\Delta_n(\beta) - \tilde{\Delta}_n(\beta)| = o_P(1).$$

The smoothness of the model now implies

$$\sup_{\beta\in B} |\tilde{\Delta}_n(\beta) - \Delta_0(\beta)| = o_P(1),$$

and thus

$$\sup_{\beta\in B} |\Delta_n(\beta) - \Delta_0(\beta)| = o_P(1).$$

As a consequence, we have for any sequence $\tilde{\beta}_n \in B$, that $\Delta_n(\tilde{\beta}_n) = o_P(1)$ implies $\Delta_0(\tilde{\beta}_n) = o_P(1)$. Hence $\tilde{\beta}_n \xrightarrow{\mathrm{P}} \beta_0$ by model identifiability, and (19.17) follows.

## 22.2 Efficient Linear Regression

This is a continuation of the linear regression example of Section 4.1.1. The model is $Y = \beta_0' Z + e$, where $Y, e \in \mathbb{R}$, $Z \in \mathbb{R}^k$ lies in a compact set $P$-almost surely, and the regression parameter $\beta_0$ lies in an known, open compact subset $B \subset \mathbb{R}^k$. We assume that $e$ has mean zero and unknown variance $\sigma^2 < \infty$ and that $e$ and $Z$ are independent, with $P[ZZ']$ being positive definite. An observation from this model consists of the pair $X = (Y, Z)$. We mentioned in Section 4.1.1 that efficient estimation for $\beta_0$ is tricky in this situation. The goal of this section is to keep the promise we made in Section 4.1.1 by developing and validating a method of efficient estimation for $\beta_0$. We will also derive a consistent estimator of the efficient Fisher information for $\beta_0$ as well as develop efficient inference for the residual distribution.

Recall that $P$ represents the true distribution and that $\eta_0$ is the density function for $e$. We need to assume that $\eta_0$ has support on all of $\mathbb{R}$ and is three time continuously differentiable with $\eta_0$, $\dot{\eta}_0$, $\ddot{\eta}_0$ and $\eta_0^{(3)}$ all uniformly bounded on $\mathbb{R}$. We also need to assume that there exist constants $0 \leq a_1, a_2 < \infty$ and $1 \leq b_1, b_2 < \infty$ such that

$$\frac{\left|\frac{\dot{\eta}_0}{\eta_0}(x)\right|}{1+|x|^{a_1}} \leq b_1, \quad \text{and} \quad \frac{\left|\frac{\ddot{\eta}_0}{\eta_0}(x)\right|}{1+|x|^{a_2}} \leq b_2,$$

for all $x \in \mathbb{R}$, and $\int_{\mathbb{R}} |x|^{(8a_1)\vee(4a_2)+6}\eta(x)dx < \infty$. The reason for allowing these functions to grow at a polynomial rate is that at least one very common possibility for $\eta$, the Gaussian density, has $\dot{\eta}/\eta$ growing at such a rate. In Exercise 22.6.2, this rate is verified, with $a_1 = 1$, $a_2 = 2$, and $b_1 = b_2 = 1$. Hence it seems unrealistic to require stronger conditions such as boundedness.

*Method of estimation.*

Our basic approach for obtaining efficient estimation will be to estimate the efficient score $\tilde{\ell}_{\beta,\eta}(x) = -(\dot{\eta}/\eta)(e)(Z-\mu) + e\mu/\sigma^2$, where $\mu \equiv PZ$, with an estimated efficient score function $\hat{\ell}_{\beta,n}$ that satisfies the conditions of Theorem 3.1. Recall that the data $(Y,Z)$ and $(e,Z)$ are equivalent for probabilistic analysis even though $e$ cannot be observed directly. The final assumptions we make are that $\tilde{I}_0 \equiv P[\tilde{\ell}_{\beta_0,\eta_0}\tilde{\ell}'_{\beta_0,\eta_0}]$ is positive definite and that the model is identifiable, in the sense that the map $\beta \mapsto \tilde{\ell}_{\beta,\eta_0}$ is continuous and has a unique zero at $\beta = \beta_0$. This identifiability can be shown to hold if the map $x \mapsto (\dot{\eta}_0/\eta_0)(x)$ is strictly monotone (see Exercise 22.6.3). Identifiability will be needed for establishing consistency of the Z-estimator $\tilde{\beta}_n$ based on the estimating equation $\beta \mapsto \mathbb{P}_n\hat{\ell}_{\beta,n}$ which will be defined presently.

The first step is to perform ordinary least-squares estimation to obtain the $\sqrt{n}$-consistent estimator $\hat{\beta}$ and then compute $\hat{F}(t) \equiv \mathbb{P}_n 1\{Y - \hat{\beta}'Z \le t\}$. As verified in Section 4.1.1, $\|\hat{F} - F\|_\infty = O_P(n^{-1/2})$, where $F$ is the cumulative distribution function corresponding to the residual density function $\eta_0$. The second step is to utilize $\hat{F}$ to estimate both $\eta_0$ and $\dot{\eta}_0$. We will utilize kernel estimators to accomplish this.

For a given sequence $h_n$ of bandwidths, define the kernel density estimator

$$\hat{\eta}_n(t) \equiv \int_{\mathbb{R}} \frac{1}{h_n}\phi\left(\frac{t-u}{h_n}\right) d\hat{F}(u),$$

where $\phi$ is the standard normal density. Taking the derivative of $t \mapsto \hat{\eta}_n(t)$ twice, we can obtain estimators for $\dot{\eta}_0$:

$$\hat{\eta}_n^{(1)}(t) \equiv -\int_{\mathbb{R}} \frac{1}{h_n^2}\left(\frac{t-u}{h_n}\right)\phi\left(\frac{t-u}{h_n}\right) d\hat{F}(u),$$

and for $\ddot{\eta}_0$:

$$\hat{\eta}_n^{(2)}(t) \equiv \int_{\mathbb{R}} \frac{1}{h_n^3}\left[\left(\frac{t-u}{h_n}\right)^2 - 1\right]\phi\left(\frac{t-u}{h_n}\right) d\hat{F}(u).$$

Define $D_n \equiv \hat{F} - F$ and use integration by parts to obtain

$$\int_{\mathbb{R}} \frac{1}{h_n}\phi\left(\frac{t-u}{h_n}\right) dD_n(u) = -\int_{\mathbb{R}} D_n(u)\frac{1}{h_n^2}\left(\frac{t-u}{h_n}\right)\phi\left(\frac{t-u}{h_n}\right) du.$$

Thus

$$\begin{aligned}\left|\int_{\mathbb{R}} \frac{1}{h_n}\phi\left(\frac{t-u}{h_n}\right) dD_n(u)\right| &\le O_P(n^{-1/2})\int_{\mathbb{R}}\left|\frac{t-u}{h_n}\right|\phi\left(\frac{t-u}{h_n}\right)h_n^{-2}du \\ &= O_P(n^{-1/2}h_n^{-1}),\end{aligned}$$

where the error terms are uniform in $t$. Moreover, since $\ddot{\eta}_0(t)$ is uniformly bounded,

$$\begin{aligned}\int_{\mathbb{R}} \frac{1}{h_n}\phi\left(\frac{t-u}{h_n}\right)[\eta_0(u)-\eta_0(t)]\,du \\ = \int_{\mathbb{R}} \frac{1}{h_n}\phi\left(\frac{t-u}{h_n}\right)\left[\dot{\eta}_0(t)(u-t)+\frac{\ddot{\eta}_0(t^*)}{2}(u-t)^2\right]du \\ = O(h_n^2),\end{aligned}$$

where $t^*$ is in the interval $[t,u]$. Thus $\|\hat{\eta}_n-\eta_0\|_\infty = O_P(n^{-1/2}h_n^{-1}+h_n^2)$.

Similar analyses (see Exercise 22.6.4) can be used to show that both $\|\hat{\eta}_n^{(1)}-\dot{\eta}_0\|_\infty = O_P(n^{-1/2}h_n^{-2}+h_n^2)$ and $\|\hat{\eta}_n^{(2)}-\ddot{\eta}_0\|_\infty = O_P(n^{-1/2}h_n^{-3}+h_n)$. Accordingly, we will assume hereafter that $h_n = o_P(1)$ and $h_n^{-1} = o_P(n^{1/6})$ so that all of these uniform errors converge to zero in probability. Now let $r_n$ be a positive sequence converging to zero with $r_n^{-1} = O_P(n^{1/2}h_n^3 \wedge h_n^{-1})$, and define $A_n \equiv \{t : \hat{\eta}_n(t) \geq r_n\}$. The third step is to estimate $t \mapsto (\dot{\eta}_0/\eta_0)(t)$ with

$$\hat{k}_n(t) \equiv \frac{\hat{\eta}_n^{(1)}(t)}{\left(\frac{r_n}{2}\vee\hat{\eta}_n(t)\right)}1\{t\in A_n\}.$$

We will now derive some important properties of $\hat{k}_n$. First note that

$$\sup_{t\in A_n}\left|\frac{\eta_0(t)}{\hat{\eta}_n(t)}-1\right| = O_P(r_n^{-1}\|\hat{\eta}_n-\eta_0\|_\infty) = O_P(r_n^{-1}[n^{-1/2}h_n^{-1}+h_n^2]).$$

Thus, for all $t\in A_n$,

$$\begin{aligned}\frac{\hat{\eta}_n^{(1)}(t)}{\hat{\eta}_n(t)} &= \frac{\dot{\eta}_0(t)}{\eta_0(t)} + O_P(r_n^{-1}\|\hat{\eta}_n^{(1)}-\dot{\eta}_0\|_\infty) \\ &\quad + \frac{\dot{\eta}_0(t)}{\eta_0(t)}O_P(r_n^{-1}\|\hat{\eta}_n-\eta_0\|_\infty) \\ &= \frac{\dot{\eta}_0(t)}{\eta_0(t)}(1+o_P(1))+o_P(1),\end{aligned}$$

where the errors are uniform over $t\in A_n$. Since $\hat{k}_n(t)$ is zero off of $A_n$, we have that

$$\frac{\hat{k}_n(t)}{1+|t|^{a_1}} \leq b_1(1+o_P(1))+o_P(1), \tag{22.6}$$

for all $t\in\mathbb{R}$, where the errors are uniform. Note also that since the support of $\eta_0$ is all of $\mathbb{R}$, we have that for every compact set $A\subset\mathbb{R}$, $A\subset A_n$ with probability going to 1 as $n\to\infty$. Thus $\|\hat{k}_n-\dot{\eta}_0/\eta_0\|_K \xrightarrow{\mathrm{P}} 0$ for every compact $K\subset\mathbb{R}$.

Second, note that the derivative of $t\mapsto\hat{k}_n(t)$, for all $t\in A_n$, is

$$\begin{aligned}
\hat{k}_n^{(1)}(t) &= \frac{\hat{\eta}_n^{(2)}(t)}{\hat{\eta}_n(t)} - \left(\frac{\hat{\eta}_n^{(1)}(t)}{\hat{\eta}_n(t)}\right)^2 \\
&= \frac{\ddot{\eta}_0(t)}{\eta_0(t)} + O_P(r_n^{-1}\|\hat{\eta}_n^{(2)} - \ddot{\eta}_0\|_\infty) \\
&\quad + \frac{\ddot{\eta}_0(t)}{\eta_0(t)} O_P(r_n^{-1}\|\hat{\eta}_n - \eta_0\|_\infty) \\
&\quad + \left(\frac{\dot{\eta}_0(t)}{\eta_0(t)}\right)^2 (1 + o_P(1)) + o_P(1) \\
&= \frac{\ddot{\eta}_0(t)}{\eta_0(t)} + O_P(r_n^{-1}[n^{-1/2}h_n^{-3} + h_n]) \\
&\quad + \frac{\ddot{\eta}_0(t)}{\eta_0(t)} O_P(r_n^{-1}[n^{-1/2}h_n^{-1} + h_n^2]) \\
&\quad + \left(\frac{\dot{\eta}_0(t)}{\eta_0(t)}\right)^2 (1 + o_P(1)) + o_P(1) \\
&= \frac{\ddot{\eta}_0(t)}{\eta_0(t)}(1 + o_P(1)) + O_P(1) \\
&\quad + \left(\frac{\dot{\eta}_0(t)}{\eta_0(t)}\right)^2 (1 + o_P(1)) + o_P(1),
\end{aligned}$$

where all error terms are uniform in $t$. Hence, since $\hat{k}_n^{(1)}(t)$ is zero for all $t \notin A_n$, we have

$$\frac{\hat{k}_n^{(1)}(t)}{(1 + |t|^{a1})^2(1 + |t|^{a_2})} \leq b_1^2(1 + o_P(1)) + b_2(1 + o_P(1)) + O_P(1),$$

where the error terms are uniform in $t$. This means that for some sequence $M_n$ which is bounded in probability as $n \to \infty$,

$$\frac{\hat{k}_n^{(1)}(t)}{1 + |t|^{(2a_1)\vee a_2}} \leq M_n, \tag{22.7}$$

for all $t \in \mathbb{R}$ and all $n$ large enough.

Third, let

$$\hat{\ell}_{\beta,n}(X) \equiv -\hat{k}_n(Y - \beta' Z)(Z - \hat{\mu}) + (Y - \beta' Z)\frac{\hat{\mu}}{\hat{\sigma}^2},$$

where $\hat{\mu} \equiv \mathbb{P}_n Z$ and $\hat{\sigma}^2 \equiv \mathbb{P}_n(Y - \hat{\beta}' Z)^2$. Our approach to estimation will be to find a zero of $\beta \mapsto \mathbb{P}_n \hat{\ell}_{\beta,n}$, i.e., a value of $\beta$ that minimizes $\|\mathbb{P}_n \hat{\ell}_{\beta,n}\|$ over $B$. Let $\check{\beta}_n$ be such a zero. We first need to show that $\check{\beta}_n$ is consistent. Then we will utilize Theorem 3.1 to establish efficiency as promised.

*Consistency.*

Combining (22.6) and (22.7) with the theorem and corollary given below, the proofs of which are given in Section 22.4, we obtain that $\{\hat{k}_n(Y-\beta'Z) : \beta \in B\}$ is contained in a $P$-Donsker class with probability going to 1 as $n \to \infty$. This implies that both $\{\hat{\ell}_{\beta,n} : \beta \in B\}$ and $\{\tilde{\ell}_{\beta,\eta_0} : \beta \in B\}$ are also contained in a $P$-Donsker class with probability going to 1 as $n \to \infty$. Thus

$$\sup_{\beta \in B} \left\| (\mathbb{P}_n - P)\left(\hat{\ell}_{\beta,n} - \tilde{\ell}_{\beta,\eta_0}\right) \right\| = O_P(n^{-1/2}).$$

Using $k_0 \equiv \dot{\eta}_0/\eta_0$, we also have that

$$\begin{aligned} (22.8)\ P\left\|\hat{\ell}_{\beta,n} - \tilde{\ell}_{\beta,\eta_0}\right\| &\leq P\left|(\hat{k}_n - k_0)(Y-\beta'Z)(Z-\hat{\mu})\right| \\ &\quad + P\left|k_0(Y-\beta'Z)(\hat{\mu}-PZ)\right| \\ &\quad + P\left|(Y-\beta'Z)\left(\frac{\hat{\mu}}{\hat{\sigma}^2} - \frac{PZ}{\sigma^2}\right)\right| \\ &\leq P\left(1\{Y-\beta'Z \notin A_n\}\left|(Y-\beta'Z)(Z-\hat{\mu})\right|\right) \\ &\quad + P\left((1+|Y-\beta'Z|^{a_1})b_1|\hat{\mu}-PZ|\right) + o_P(1) \\ &\leq o_P(1), \end{aligned}$$

as a consequence of previously determined properties of $\hat{k}_n$ and $k_0$. Hence

$$\sup_{\beta \in B} \left\| \mathbb{P}_n \hat{\ell}_{\beta,n} - P\tilde{\ell}_{\beta,\eta_0} \right\| = o_P(1).$$

Thus the identifiability of the model implies that $\check{\beta}_n \xrightarrow{\mathrm{P}} \beta_0$.

THEOREM 22.1 *Let $\mathcal{F}$ be a class of differentiable functions $f : \mathbb{R} \mapsto \mathbb{R}$ such that for some $0 \leq \alpha, \beta < \infty$ and $1 \leq M < \infty$,*

$$\sup_{f \in \mathcal{F}} \sup_{x \in \mathbb{R}} \left| \frac{f(x)}{1+|x|^\alpha} \right| \leq M \quad \text{and} \quad \sup_{f \in \mathcal{F}} \sup_{x \in \mathbb{R}} \left| \frac{\dot{f}(x)}{1+|x|^\beta} \right| \leq M.$$

*Then for some constant $K$ depending only on $\alpha, \beta, M$, we have*

$$\log N_{[]}(\delta, \mathcal{F}, L_2(Q)) \leq \frac{K\left[1+Q|X|^{4(\alpha \vee \beta)+6}\right]^{1/4}}{\delta},$$

*for all $\delta > 0$ and all probability measures $Q$.*

COROLLARY 22.2 *Let $\mathcal{F}$ be as defined in Theorem 22.1, and define the class of functions*

$$\mathcal{H} \equiv \{f(Y-\beta'Z) : \beta \in B,\ f \in \mathcal{F}\},$$

*where $B \subset \mathbb{R}^k$ is bounded. Then, for every probability measure $P$ for which $P|Y-\beta_0'Z|^{4(\alpha\vee\beta)+6} < \infty$, for some $\beta_0 \in B$, and for which $Z$ lies in a bounded subset of $\mathbb{R}^k$ $P$-almost surely, $\mathcal{H}$ is $P$-Donsker.*

*Efficiency.*

Provided we can establish that

$$P\left[k_0^2(Y-\check{\beta}_n'Z)1\{Y-\check{\beta}_n'Z\notin A_n\}\right] \quad = \quad o_P(1), \tag{22.9}$$

the result in (22.8) can be strengthened to conclude that Condition (3.7) holds. Accordingly, the left-hand-side of (22.9) is bounded by

$$P[(1+|Y-\check{\beta}_n'Z|^{a_1})^2 b_1^2 1\{Y-\check{\beta}_n'Z\notin K_n\}+o_P(1),$$

for some sequence of increasing compact sets $K_n$, by previously established properties of $k_0$ and $A_n$. Since $Y=\beta_0'Z+e$, we obtain that $|Y-\check{\beta}_n'Z|^{a_1}\lesssim |e\vee 1|^{a_1}$. Hence, for some sequence $t_n\to\infty$, the left-hand-size of (22.9) is bounded up to a constant by $P\left[|e|^{2a_1}1\{|e|>t_n\}\right]=o(1)$. Thus (3.7) holds.

The only condition of Theorem 3.1 remaining to be established is (3.6). Consider the distribution $P_{\check{\beta}_n,\eta_0}$. Under this distribution, the mean of $\hat{e}\equiv Y-\check{\beta}_n'Z$ is zero and $\hat{e}$ and $Z$ are independent. Hence

$$P_{\check{\beta}_n,\eta_0}\hat{\ell}_{\check{\beta}_n,n}=P_{\check{\beta}_n,\eta_0}\left[-\hat{k}_n(Y-\check{\beta}_n'Z)(PZ-\hat{\mu})\right].$$

However, $\hat{\mu}-PZ=O_P(n^{-1/2})$ and

$$P_{\check{\beta}_n,\eta_0}\hat{k}_n(Y-\check{\beta}_n'Z)=o_P(1)+P\left[k_0(Y-\check{\beta}_n'Z)1\{Y-\check{\beta}_n'Z\notin A_n\}\right].$$

Previous arguments can be recycled to established that the last term on the right is $o_P(1)$. Thus $P_{\check{\beta}_n,\eta_0}\hat{\ell}_{\check{\beta}_n,n}=o_P(n^{-1/2})$, and (3.6) follows. This means that all of the conditions of Theorem 3.1 are established ... except the requirement that $\mathbb{P}_n\hat{\ell}_{\check{\beta}_n,n}=o_P(n^{-1/2})$. The issue is that even though we are assuming $\|\mathbb{P}_n\hat{\ell}_{\beta,n}\|$ is minimized at $\beta=\check{\beta}_n$, there is no guarantee at this point that such a minimum is of the order $o_P(n^{-1/2})$. Once we verify this, however, the theorem will finally yield that $\check{\beta}_n$ is efficient.

To accomplish this last step, we will use arguments in the proof of Theorem 3.1, which proof is given in Section 19.1, to show that if $\tilde{\beta}_n$ satisfies $\mathbb{P}_n\tilde{\ell}_{\tilde{\beta}_n,\eta_0}=o_P(n^{-1/2})$ then $\mathbb{P}_n\hat{\ell}_{\tilde{\beta}_n,n}=o_P(n^{-1/2})$. The existence of such a $\tilde{\beta}_n$ is guaranteed by the local quadratic structure of $\beta\mapsto\mathbb{P}_n\tilde{\ell}_{\beta,\eta_0}$ and the assumed positive-definiteness of $\tilde{I}_0$. Now,

$$\begin{aligned}\mathbb{P}_n(\hat{\ell}_{\tilde{\beta}_n,n}-\tilde{\ell}_{\tilde{\beta}_n,\eta_0}) &= n^{-1/2}\mathbb{G}_n(\hat{\ell}_{\tilde{\beta}_n,n}-\tilde{\ell}_{\tilde{\beta}_n,\eta_0})+P_{\tilde{\beta}_n,\eta_0}(\hat{\ell}_{\tilde{\beta}_n,n}-\tilde{\ell}_{\tilde{\beta}_n,\eta_0})\\ &\quad -(P_{\tilde{\beta}_n,\eta_0}-P)(\hat{\ell}_{\tilde{\beta}_n,n}-\tilde{\ell}_{\tilde{\beta}_n,\eta_0})\\ &\equiv E_n+F_n-G_n.\end{aligned}$$

By the $P$-Donsker properties of $\hat{\ell}_{\tilde{\beta}_n,n}$ and $\tilde{\ell}_{\tilde{\beta}_n,\eta_0}$ combined with the fact that $P\left(\hat{\ell}_{\tilde{\beta}_n,n}-\tilde{\ell}_{\tilde{\beta}_n,\eta_0}\right)^2=o_P(1)$, we have that $E_n=o_P(n^{-1/2})$. Condition (3.6) implies that $F_n=o_P(n^{-1/2}+\|\tilde{\beta}_n-\beta_0\|)=o_P(n^{-1/2})$. The

last equality follows from the fact that $\tilde{\beta}_n$ is efficient (since we are using the exact efficient score for inference). Finally, arguments used in the second paragraph of the proof of Theorem 3.1 yield that $G_n = o_P(n^{-1/2})$. Thus $o_P(n^{-1/2}) = \|\mathbb{P}_n \hat{\ell}_{\tilde{\beta}_n,n}\| \geq \|\mathbb{P}_n \hat{\ell}_{\check{\beta}_n,n}\|$, and the desired efficiency of $\check{\beta}_n$ follows.

*Variance estimation for $\check{\beta}_n$.*

The next issue for this example is to obtain a consistent estimator of $\tilde{I}_0 = P[\tilde{\ell}_{\beta_0,\eta_0}\tilde{\ell}'_{\beta_0,\eta_0}]$, which estimator we denote $\hat{I}_n$. This will facilitate inference for $\beta_0$, since we would then have that $n(\check{\beta}_n - \beta_0)'\hat{I}_n(\check{\beta}_n - \beta_0)$ converges weakly to a chi-square random variable with $k$ degrees of freedom. We now show that $\hat{I}_n \equiv \mathbb{P}_n \left[\hat{\ell}_{\check{\beta}_n,n}\hat{\ell}'_{\check{\beta}_n,n}\right]$ is such an estimator.

Since $\hat{\ell}_{\check{\beta}_n,n}$ is contained in a $P$-Donsker class with probability tending to 1 for $n \to \infty$, it is also similarly contained in a $P$-Glivenko-Cantelli class. Recycling previous arguments, we can verify that

$$P\left[\hat{\ell}_{\check{\beta}_n,n}\hat{\ell}'_{\check{\beta}_n,n}\right] = P\left[\tilde{\ell}_{\beta_0,\eta_0}\tilde{\ell}'_{\beta_0,\eta_0}\right] + o_P(1).$$

Since the product of two $P$-Glivenko-Cantelli classes is also $P$-Glivenko-Cantelli provided the product of the two envelopes has bounded expectation, we now have that $(\mathbb{P}_n - P)\left[\hat{\ell}_{\check{\beta}_n,n}\hat{\ell}'_{\check{\beta}_n,n}\right] = o_P(1)$, after arguing along subsequences if needed. The desired consistency follows.

*Inference for the residual distribution.*

It is not hard to show that

$$\check{F}_n(t) \equiv \mathbb{P}_n 1\{Y - \check{\beta}'_n Z \leq t\}$$

is efficient for estimating the true residual distribution $F_0(t)$, uniformly over $t \in \mathbb{R}$, by using arguments in Section 4.1.1 and concepts in Chapters 3 and 18. The verification of this is saved as an exercise (see Exercise 22.6.5). This means that $(\check{\beta}_n, \check{F}_n)$ are jointly efficient for $(\beta_0, F_0)$.

Moreover, valid joint inference for both parameters can be obtained by using the piggyback bootstrap $(\beta_n, \check{F}^\circ_n)$, where $\beta_n = \check{\beta}_n + n^{-1/2}\hat{I}_n^{-1/2}U$ and $\check{F}^\circ_n(t) = \mathbb{P}^\circ_n 1\{Y - \beta'_n Z \leq t\}$, $U$ is a standard $k$-variate normal deviate independent of the data, and $\mathbb{P}^\circ_n$ is the weighted empirical measure with random weights $\xi_1, \ldots, \xi_n$ which are independent of both $U$ and the data and which satisfy the conditions given in Section 20.2.2. The verification of this is also saved as an exercise (see Exercise 22.6.6).

## 22.3 Temporal Process Regression

In this section, we will study estimation for a varying coefficient regression model for temporal process data. The material is adapted from Fine,

Yan and Kosorok (2004). Consider, for example, bone marrow transplantation studies in which the time-varying effect of a certain medication on the prevalence of graft versus host (GVH) disease may be of scientific interest. Let the outcome measure—or "response"—be denoted $Y(t)$, where $t$ is restricted to a finite time interval $[l, u]$. In the example, $Y(t)$ is a dichotomous process indicating the presence of GVH at time $t$. More generally, we allow $Y$ to be a stochastic process, but we require $Y$ to have square-integrable total variation over $[l, u]$.

We model the mean of the response $Y$ at time $t$ as a function of a $k$-vector of time-dependent covariates $X(t)$ and a time-dependent stratification indicator $S(t)$ as follows:

$$\mathrm{E}[Y(t)|X(t), S(t) = 1] \quad = \quad g^{-1}(\beta'(t)X(t)), \tag{22.10}$$

where the link function $g$ is monotone, differentiable and invertible, and $\beta(t) = \{\beta_1(t), \ldots, \beta_k(t)\}'$ is a $k$-vector of time-dependent regression coefficients. A discussion of the interpretation and uses of this model is given in Fine, Yan and Kosorok. For the bone marrow example, $g^{-1}(u) = e^u/(1+e^u)$ would yield a time-indexed logistic model, with $\beta(t)$ denoting the changes in log odds ratios over time for GVH disease prevalence per unit increase in the covariates.

Note that no Markov assumption is involved here since the conditioning in (22.10) only involves the current time and not previous times. In addition to the stratification indicator $S(t)$, it is useful to include a non-missing indicator $R(t)$, for which $R(t) = 1$ if $\{Y(t), X(t), S(t)\}$ is fully observed at $t$, and $R(t) = 0$ otherwise. We assume that $Y(t)$ and $R(t)$ are independent conditionally on $\{X(t), S(t) = 1\}$. The data structure is similar in spirit to that described in Nadeau and Lawless (1998). The approach we take for inference is to utilize that fact that the model only posits the conditional mean of $Y(t)$ and not the correlation structure. Thus we can construct "working independence" estimating equations to obtain simple, nonparametric estimators. The pointwise properties of these estimators follows from standard estimating equation results, but uniform properties are quite nontrivial to establish since martingale theory is not applicable here. We will use empirical process methods.

The observed data consists of $n$ independent and identically distributed copies of $\{R(t) : t \in [l, u]\}$ and $(\{Y(t), X(t), S(t)\} : R(t) = 1, t \in [l, u])$. We can compute an estimator $\hat{\beta}_n(t)$ for each $t \in [l, u]$ as the root of $U_n(\beta(t), t) \equiv \mathbb{P}_n A(\beta(t), t)$, where

$$A(\beta(t), t) \equiv S(t)R(t)D(\beta(t))V(\beta(t), t)\left[Y(t) - g^{-1}(\beta'(t)X(t))\right],$$

where $D(u) \equiv \partial[g^{-1}(u'X(t))]/(\partial u)$ is a column $k$-vector-valued function and $V(\beta(t), t)$ is a time-dependent and possibly data-dependent scalar-valued weight function. Here are the specific data and estimating equation assumptions we will need:

(i) $(S_i, R_i, X_i, Y_i)$, $i = 1, \ldots, n$, are i.i.d. and all component processes are cadlag. We require $S$, $R$ and $X$ to all have total variation over $[l, u]$ bounded by a fixed constant $c < \infty$, and we require $Y$ to have total variation $\tilde{Y}$ over $[l, u]$ which has finite second moment.

(ii) $t \mapsto \beta(t)$ is cadlag on $[l, u]$.

(iii) $h \equiv g^{-1}$ and $\dot{h} = \partial h(u)/(\partial u)$ are Lipschitz continuous and bounded above and below on compact sets.

(iv) We require

$$\inf_{t \in [l,u]} \text{eigmin}\, P\left[S(t)R(t)X(t)X'(t)\right] > 0,$$

where eigmin denotes the minimum eigenvalue of a matrix.

(v) For all bounded $B \subset \mathbb{R}^k$, the class of random functions $\{V(b,t) : b \in B, t \in [l,u]\}$ is bounded above and below by positive constants and is BUEI and PM (recall these definitions from Sections 9.1.2 and 8.2).

Note that Condition (v) is satisfied if $V(b,t) \equiv v(b'X(t))$, where $u \mapsto v(u)$ is a nonrandom, positive and Lipschitz continuous function.

The form of the estimator will depend on the form of the observed data. The estimator jumps at those $M$ times where

$$t \mapsto (\{Y_i(t), X_i(t), S_i(t)\} : R_i(t) = 1)$$

and $t \mapsto R_i(t)$ jumps, $i = 1, \ldots, n$. If $Y_i(t)$ and $X_i(t)$ are piecewise constant, then so also is $\hat{\beta}_n$. In this situation, finding $\hat{\beta}_n$ (as a process) involves solving $t \mapsto U_n(\beta(t), t)$ at these $M$ time points. For most practical applications, $Y$ and $X$ will be either piecewise-constant or continuous, and, therefore, so will $\hat{\beta}_n$. In the piecewise-constant case, we can interpolate in a right-continuous manner between the $M$ jump points, otherwise, we can smoothly interpolate between them. When $M$ is large, the differences between these two approaches will be small. The bounded total variation assumptions on the data make the transition from pointwise to uniform estimation and inference both theoretically possible and practically feasible. In this light, we will assume hereafter that $\hat{\beta}_n$ can be computed at every value of $t \in [l, u]$.

We will now discuss consistency, asymptotic normality, and inference based on simultaneous confidence bands. We will conclude this example with some comments on optimality of the estimators in this setting. Several interesting examples of data analyses and simulation studies for this set-up are given in Fine, Yan and Kosorok (2004).

*Consistency.*

The following theorem gives us existence and consistency of $\hat{\beta}_n$ and the above conditions:

THEOREM 22.3 *Assume (22.10) holds with true parameter* $\{\beta_0(t) : t \in [l,u]\}$, *where* $\sup_{t\in[l,u]} |\beta_0(t)| < \infty$. *Let* $\hat{\beta}_n = \{\hat{\beta}_n(t) : t \in [l,u]\}$ *be the smallest, in uniform norm, root of* $\{U_n(\beta(t),t) = 0 : t \in [l,u]\}$. *Then such a root exists for all n large enough almost surely, and*

$$\sup_{t\in[l,u]} \left|\hat{\beta}_n(t) - \beta_0(t)\right| \overset{\text{as*}}{\to} 0.$$

**Proof.** Define

$$C(\gamma,\beta,t) \equiv S(t)R(t)D(\gamma(t))V(\gamma,t)\left[Y(t) - h(\beta'(t)X(t))\right],$$

where $\gamma,\beta \in \{\ell_c^\infty([l,u])\}^k$ and $\ell_c^\infty(H)$ is the collection of bounded real functions on the set $H$ with absolute value $\le c$ (when $c = \infty$, the subscript $c$ is omitted and the standard definition of $\ell^\infty$ prevails). Let

$$\mathcal{G} \equiv \left\{C(\gamma,\beta,t) : \gamma,\beta \in \{\ell_c^\infty([l,u])\}^k, t \in [l,u]\right\}.$$

The first step is to show that $\mathcal{G}$ is BUEI and PM with square-integrable envelope for each $c < \infty$. This implies that $\mathcal{G}$ is $P$-Donsker and hence also $P$-Glivenko-Cantelli. We begin by observing that the classes

$$\{\beta'(t)X(t) : \beta \in \{\ell_c^\infty([l,u])\}^k, t \in [l,u]\} \text{ and } \{b'X(t) : b \in [-c,c]^k, t \in [l,u]\}$$

are equivalent. Next, the following lemma yields that processes with square-integrable total variation are BUEI and PM (the proof is given below):

LEMMA 22.4 *Let* $\{W(t) : t \in [l,u]\}$ *be a cadlag stochastic process with square-integrable total variation* $\tilde{W}$, *then* $\{W(t) : t \in [l,u]\}$ *is BUEI and PM with envelop* $2\tilde{W}$.

**Proof.** Let $\mathcal{W}$ consist of all cadlag functions $w : [l,u] \mapsto \mathbb{R}$ of bounded total variation and let $\mathcal{W}_0$ be the subset consisting of monotone increasing functions. Then the functions $w_j : \mathcal{W} \mapsto \mathcal{W}_0$, where $w_1(w)$ extracts the monotone increasing part of $w$ and $w_2(w)$ extracts the negative of the monotone decreasing part of $w$, are both measurable. Moreover, for any $w \in \mathcal{W}$, $w = w_1(w) - w_2(w)$. Lemma 9.10 tells us that $\{w_j(W)\}$ is VC with index 2, for both $j = 1,2$. It is not difficult to verify that cadlag monotone increasing processes are PM (see Exercise 22.6.7). Hence we can apply Part (iv) of Lemma 9.17 to obtain the desired result.□

By applying Lemma 9.17, we obtain that $\mathcal{G}$ is BUEI and PM with square-integrable envelope and hence is $P$-Donsker by Theorem 8.19 and the statements immediately following the theorem.

The second step is to use this Donsker property to obtain existence and consistency. Accordingly, we now have for each $c < \infty$ and all $\tilde{\beta} \in \{\ell_c^\infty([l,u])\}^k$, that

$$\begin{aligned} & U_n(\tilde{\beta}(t),t) \\ &= \mathbb{P}_n \Big\{ C(\tilde{\beta},\beta_0,t) - S(t)R(t)D(\tilde{\beta}(t))V(\tilde{\beta}(t),t) \\ &\quad \times \Big[ h(\tilde{\beta}'(t)X(t)) - h(\beta_0'(t)X(t)) \Big] \Big\} \\ &= -\mathbb{P}_n \Big[ S(t)R(t)X(t)X'(t)\dot{h}(\check{\beta}'(t)X(t))\dot{h}(\tilde{\beta}'(t)X(t))V(\tilde{\beta}(t),t) \Big] \\ &\quad \times \Big\{ \tilde{\beta}(t) - \beta_0(t) \Big\} + \epsilon_n(t), \end{aligned}$$

where $\check{\beta}(t)$ is on the line segment between $\tilde{\beta}(t)$ and $\beta_0(t)$ and $\epsilon_n(t) \equiv \mathbb{P}_n C(\tilde{\beta},\beta_0,t)$, $t \in [l,u]$. Since $\mathcal{G}$ is $P$-Glivenko-Cantelli and $PC(\tilde{\beta},\beta_0,t) = 0$ for all $t \in [l,u]$ and $\tilde{\beta} \in \{\ell_c^\infty([l,u])\}^k$, we have $\sup_{t\in[l,u]} |\epsilon_n(t)| \overset{\text{as*}}{\to} 0$. By Condition (ii) and the uniform positive-definiteness assured by Condition (iv), the above results imply that $U_n(\tilde{\beta}(t),t)$ has a uniformly bounded solution $\hat{\beta}_n$ for all $n$ large enough. Hence $\|U_n(\hat{\beta}_n(t),t)\| \geq c\|\hat{\beta}_n(t) - \beta_0(t)\| - \epsilon_n^*(t)$, where $c > 0$ does not depend on $t$ and $\|\epsilon_n^*\|_\infty \overset{\text{as*}}{\to} 0$. This follows because $\{S(t)R(t)X(t)X(t)' : t \in [l,u]\}$ is $P$-Glivenko-Cantelli using previous arguments. Thus the desired uniform consistency follows.□

*Asymptotic normality.*

The following theorem establishes both asymptotic normality and an asymptotic linearity structure which we will utilize later for inference:

THEOREM 22.5 *Under the conditions of Theorem 22.3, $\hat{\beta}_n$ is asymptotically linear with influence function $\psi(t) \equiv -[H(t)]^{-1}A(\beta_0(t),t)$, where*

$$H(t) \equiv P\left[S(t)R(t)D(\beta_0(t))V(\beta_0(t),t)D'(\beta_0(t))\right],$$

*and $\sqrt{n}(\hat{\beta}_n - \beta_0)$ converges weakly in $\{\ell^\infty([l,u])\}^k$ to a tight, mean zero Gaussian process $\mathcal{X}(t)$ with covariance*

$$\Sigma(s,t) \equiv P[\mathcal{X}(s)\mathcal{X}'(t)] = P[\psi(s)\psi'(t)].$$

**Proof.** By Theorem 22.3, we have for all $n$ large enough,

$$\begin{aligned} 0 &= \sqrt{n}U_n(\hat{\beta}_n(t),t) \\ &= \sqrt{n}\mathbb{P}_n A(\beta_0(t),t) + \sqrt{n}\mathbb{P}_n \Big[ A(\hat{\beta}_n(t),t) - A(\beta_0(t),t) \Big] \\ &= \sqrt{n}\mathbb{P}_n A(\beta_0(t),t) + \sqrt{n}\mathbb{P}_n \Big[ C(\hat{\beta}_n,\beta_0,t) - C(\beta_0,\beta_0,t) \Big] \\ &\quad -\sqrt{n}\mathbb{P}_n \Big[ S(t)R(t)D(\hat{\beta}_n(t))V(\hat{\beta}_n(t),t) \\ &\quad \times \Big\{ h(\hat{\beta}_n'(t)X(t)) - h(\beta_0'(t)X(t)) \Big\} \Big] \\ &\equiv \sqrt{n}\mathbb{P}_n A(\beta_0(t),t) + J_n(t) - K_n(t). \end{aligned}$$

Since $\mathcal{G}$ is $P$-Donsker and since sums of Donsker classes are Donsker, and also since

$$\sup_{t\in[l,u]} P\left\{C(\hat{\beta}_n,\beta_0,t)-C(\beta_0,\beta_0,t)\right\}^2 \xrightarrow{\mathrm{P}} 0,$$

we have that $\sup_{t\in[l,u]}|J_n(t)| = o_P(1)$.

By previous arguments, we also have that

$$K_n(t) = [H(t)+\epsilon_n^{**}(t)]\sqrt{n}(\hat{\beta}_n(t)-\beta_0(t)),$$

where a simple extension of previous arguments yields that $\sup_{t\in[l,u]}|\epsilon_n^{**}(t)| = o_P(1)$. This now yields the desired asymptotic linearity. The weak convergence follows since $\{A(\beta_0(t),t) : t\in[l,u]\}$ is a subset of the Donsker class $\mathcal{G}$.□

*Simultaneous confidence bands.*

We now utilize the asymptotic linear structure derived in the previous theorem to develop simultaneous confidence band inference. Define

$$\hat{H}_n(t) \equiv \mathbb{P}_n\left[S(t)R(t)D(\hat{\beta}_n(t))V(\hat{\beta}_n(t),t)D'(\hat{\beta}_n(t))\right]$$

and

$$\hat{\Sigma}_n(s,t) \equiv \hat{H}_n^{-1}(s)\mathbb{P}_n\left[A(\hat{\beta}_n(s),s)A'(\hat{\beta}_n(t),t)\right]\hat{H}_n^{-1}(t),$$

and let $\hat{M}(t) = [\hat{M}_1(t),\ldots,\hat{M}_k(t)]'$ be the component-wise square root of the diagonal of $\hat{\Sigma}_n(t,t)$. Define also

$$I_n^\circ(t) \equiv n^{-1/2}\sum_{i=1}^n G_i\left[\operatorname{diag}\hat{M}(t)\right]^{-1}\left\{\hat{H}_n(t)\right\}^{-1}A_i(\hat{\beta}_n(t),t),$$

where $G_1,\ldots,G_n$ are i.i.d. standard normal deviates independent of the data. Now let $m_n^\circ(\alpha)$ be the $1-\alpha$ quantile of the conditional sampling distribution of $\sup_{t\in[l,u]}\|I_n^\circ(t)\|$.

The next theorem establishes that $\hat{\Sigma}$ is consistent for $\Sigma$ and that

$$\hat{\beta}_n(t) \quad \pm \quad n^{-1/2}m_n^\circ(\alpha)\hat{M}(t) \tag{22.11}$$

is a $1-\alpha$-level simultaneous confidence band for $\beta_0(t)$, simultaneous for all $t\in[u,l]$. A nice property of the band in (22.11) is that its size is rescaled in a time-dependent manner proportionally to the component-wise standard error. This means that time-dependent changes in precision are accurately reflected by the confidence band.

THEOREM 22.6 *Under the conditions of Theorem 22.3, $\hat{\Sigma}(s,t)$ is uniformly consistent for $\Sigma(s,t)$, over all $s,t\in[l,u]$, almost surely. If, in addition, $\inf_{t\in[l,u]}$ eigmin $\Sigma(t,t) > 0$, then the $1-\alpha$ confidence band given in (22.11) is simultaneously valid asymptotically.*

**Proof.** The proof of uniform consistency of $\hat{\Sigma}$ follows from minor modifications of previous arguments. Provided the minimum eigenvalue condition holds, $\hat{M}(t)$ will be asymptotically bounded both above and below uniformly over $t \in [l, u]$ and uniformly consistent for the component-wise square root of the diagonal of $\Sigma(t, t)$, which we denote $M_0(t)$. The arguments in Section 20.2.3 are applicable, and we can establish, again by recycling earlier arguments, that $I_n^\circ(t) \underset{G}{\overset{\mathrm{P}}{\rightsquigarrow}} M_0^{-1}(t)\mathcal{X}(t)$ in $\{\ell^\infty([l, u])\}^k$. The desired conclusions now follow.□

*Optimality.*

Note that the estimating equation $U_n$ is an infinite-dimensional analog of that in Liang and Zeger (1986), with an independence working assumption across $t$. This avoids having to worry about modeling temporal correlations. Nevertheless, incorporating weights which account for these dependencies may improve the efficiency of the estimators. With standard longitudinal data, the response's dimension is small and specifying the dependencies is not difficult (see Prentice and Zhao, 1991; Zhao, Prentice and Self, 1990). Moreover, misspecification of the dependence in this case does not bias the estimators. It is not clear whether this approach can be readily extended to temporal process responses.

In order to assess efficiency for the functional model, we need to select a Hilbert space in $D[l, u]$ and define a score operator on that space. A reasonable choice is the space, which we will denote $H_1$, that has inner product $\langle a, b\rangle_1 = \int_{[l,u]} a(s)b(s)d\mu(s)$, where $\mu$ is a bounded measure. For ease of exposition, we will assume throughout this section that $X(t) = X$ is time-independent. Now let $\check{U}_n(\beta)$ be the empirical expectation of $\hat{D}_X R S \hat{V}_{R,S,X} \dot{Y}(\beta)$, where $\hat{D}_X : H_1 \mapsto H_p$ takes $a \in H_1$ and multiplies it component-wise to obtain a vector function with components

$$a(t)\frac{\partial g^{-1}(u'X)}{\partial u}\bigg|_{u=\tilde{\beta}_n(t)},$$

for a sequence $\tilde{\beta}_n$ uniformly consistent for $\beta_0$, and where $\dot{Y}(\beta)(t) \equiv Y(t) - g^{-1}(\beta(t)'X)$. We assume that $\hat{V}_{R,S,X} : H_1 \mapsto H_1$ is an estimated operator that converges in probability. Let $\tau_{j,R,S} \equiv \{t \in [l, u] : R(t)S(t) = j\}$, for $j = 0, 1$, and restrict $\hat{V}_{R,S,X}$ such that if $g = \hat{V}_{R,S,X} f$, then $g(t) = f(t)$ for all $t \in \tau_{0,R,S}$. In other words, the operator functions as the pointwise identity for all $t$ where $S(t)R(t) = 0$. Arguments similar to those used above to establish consistency and asymptotic normality of $\hat{\beta}_n$ can be extended to verify that $\check{\beta}_n$ satisfying $\check{U}_n(\check{\beta}_n) = 0$ exists for all $n$ large enough and is consistent and, moreover, that $\sqrt{n}(\check{\beta}_n - \beta_0)$ converges weakly in $\{\ell^\infty([l, u])\}^k$ to a tight Gaussian process.

Our efforts toward achieving optimality, therefore, can essentially be reduced to finding a $\hat{V}_{R,S,X}$ that is uniformly consistent for an operator which

makes $\check{U}_n$ optimal with respect to $H_1$. What this translates to is that the variance of $\langle h, \check{\beta}_n \rangle_1$ is minimal over all estimators from equations of the form specified above. In other words, it is variance-minimizing for all $h \in H_1$, corresponding to all linear functionals on $H_1$. This is a natural extension of the optimal estimating function concept for finite-dimensional parameters (see Godambe, 1960). It is not difficult to derive that the limiting covariance operator for $\sqrt{n}(\check{\beta}_n - \beta_0)$ is

$$\begin{aligned} (22.12) \qquad \check{V} \equiv & \{P[D_X RSV_{R,S,X} RSD_X^*]\}^{-1} \\ & \times P\left[D_X RSV_{R,S,X} W_{R,S,X} V_{R,S,X}^* RSD_X^*\right] \\ & \times \{P[D_X RSV_{R,S,X} RSD_X^*]\}^{-1}, \end{aligned}$$

provided the inverses exist, where $D_X$ and $V_{R,S,X}$ are the asymptotic limits of $\hat{D}_X$ and $\hat{V}_{R,S,X}$, respectively, the superscript $*$ denotes adjoint, and $W_{R,S,X} : H_1 \mapsto H_1$ is the self-adjoint operator which takes $a \in H_1$ to

$$W_{R,S,X} a(t) \equiv \begin{cases} \int_{[l,u]} \sigma_X(s,t) a(s) R(s) S(s) d\mu(s), & \text{for all } t \in \tau_{1,R,S}, \\ a(t), & \text{for all } t \in \tau_{0,R,S}, \end{cases}$$

where

$$\sigma_x(s,t) \equiv P\left[\dot{Y}(\beta_0)(s)\dot{Y}(\beta_0)(t) \middle| X = x, R(s)S(s) = R(t)S(t) = 1\right].$$

Using Hilbert space techniques, including those presented in Chapter 17, we can adapt Godambe's (1960) proof to show that the optimal choice for $V_{R,S,X}$ is $W_{R,S,X}^{-1}$, provided the inverse exists. In this instance, the right side of (22.12) simplifies to

$$(22.13) \qquad \check{V} = \{P[D_X W_{R,X,S} D_X^*]\}^{-1}.$$

When the measure $\mu$ used in defining the inner product on $H_1$ only gives mass to a finite set of points, then (22.13) follows from the standard finite-dimensional results. Thus, one may view the optimal score operator for the functional model (22.10) as a generalization of Liang and Zeger (1986).

An important question is whether or not $W_{R,S,X}^{-1}$ exists. If not, then, although the lower bound (22.13) may exist, it may not be achievable through any estimating equation of the form $\check{U}_n$. Unlike with the finite-dimensional setting, where $W_{R,S,X}$ is a matrix, serious problems can arise in the functional set-up. Consider, for example, the case where $H_1$ is truly infinite-dimensional, as when $\mu$ is Lebesgue measure on $[l,u]$. Suppose, for simplicity, that $R(t)S(t) = 1$ almost surely for all $t \in [l,u]$ and that $\dot{Y}(\beta_0)$ is a smooth Gaussian process with conditional covariance function $\sigma_x(s,t)$ satisfying $\int_{[l,u]} \int_{[l,u]} \sigma_x^2(s,t) ds dt \leq M$ for all $x$, where $M < \infty$, and the Hilbert space is $L_2[l,u]$. Note that such smooth Gaussian processes can be obtained, for example, by integrating less smooth Gaussian processes with

bounded covariances. For any finite collection of time points in $[l, u]$, the optimum equation exists if the matrix inverse of $W_{R,S,X}$ restricted to those time points exists. However, this does not typically imply that $W_{R,S,X}^{-1}$ exists on $L_2[l, u]$.

To see the issue, note that the image of $W_{R,S,X}$ can be shown using Picard's theorem (see, for example, Wahba, 1990, Chapter 8) to be the reproducing kernel Hilbert space with reproducing kernel

$$K_x(t, v) \equiv \int_{[l,u]} \sigma_x(t, s)\sigma_x(s, v)ds.$$

This implies that $W_{R,S,X}^{-1}$ exists only on a strict subset of the reproducing kernel Hilbert space with reproducing kernel $\sigma_x$. It can be further shown that a tight Gaussian process with covariance $\sigma_x$ is, with probability 1, not a member of the reproducing kernel Hilbert space with kernel $\sigma_x$ (see Page 5 of Wahba, 1990) and hence not a member of the reproducing kernel Hilbert space with kernel $K_x$. Thus, with probability 1, $W_{R,S,X}^{-1}\dot{Y}(\beta_0)$ does not exist. Hence, even if $W_{R,S,X}^{-1}$ is valid on a subspace of $H_1$ that is large enough for $\hat{V}$ to exist, the score operator will not exist. It is easy to conjecture that similar difficulties arise with more complicated non-Gaussian data.

On the other hand, even though $W_{R,S,X}^{-1}$ may not exist, we do have, for every $\epsilon > 0$, that $\check{W}_{R,S,X} \equiv \epsilon I + W_{R,S,X}$—where $I$ is the identity operator—does have an inverse over all of $L_2[l, u]$, by standard results for Volterra integral equations (see Section 3.3 of Kress, 1999). This means that it is possible to construct a score operator, using $\check{W}_{R,S,X}$, with deterministic, small $\epsilon > 0$, that is arbitrarily close to the optimal estimating equation. In practical settings, the effectiveness of $\check{W}$ will depend on the choice of $\epsilon$ and the degree of temporal correlation. The stronger the correlations, the more unstable the inverse. While optimality may be out of reach, significant improvements over the estimating equation $U_n$ are possible by utilizing the correlation structure of $\dot{Y}(\beta_0)$ as we have described above. Further discussions of these concepts can be found in Fine, Yan and Kosorok (2004).

An important conclusion from this example is that efforts to achieve optimality may yield meaningful gains in efficiency even in those settings where optimality is unachievable.

## 22.4 A Partly Linear Model for Repeated Measures

In this section, we study a marginal partly linear regression model for repeated measures data. We assume the data are i.i.d. observations $X_1, \ldots,$ $X_n$ with $0 < q < \infty$ repeated measures per individual $X_i$. More precisely, an individual observation is $X_i \equiv (Y_i, U_i)$, where $Y_i = (Y_{i1}, \ldots, Y_{iq})'$ is an $q$-vector of possibly dependent outcome measures, $U_i \equiv (Z_i, W_i)$ is a matrix

of regression covariates, $Z_i \equiv (Z_{i1}, \ldots, Z_{iq})'$ with $Z_{ij} \equiv (Z_{ij1}, \ldots, Z_{ijp})' \in \mathbb{R}^p$, $j = 1, \ldots, q$, and $W_i \equiv (W_{i1}, \ldots, W_{iq})' \in [0,1]^m$. Consistent with these definitions, we also define $U_{ij} \equiv (Z'_{ij}, W_{ij})'$, so that $U_i = (U_{i1}, \ldots, U_{iq})'$. This kind of data arises in longitudinal studies, family studies, and in many other settings involving clustered data. In many applications, $q$ may actually vary from individual to individual, but, for ease of exposition, we assume in this section that $q$ is fixed throughout.

The model we assume is $Y_{ij} = \beta_0' Z_{ij} + h_0(W_{ij}) + \epsilon_{ij}$, $j = 1, \ldots, q$, where $\beta \in \mathbb{R}^p$, $h_0 \in \mathcal{H}$, $\mathcal{H}$ is the space of functions $h : [0,1] \mapsto \mathbb{R}$ with $J(h) < \infty$, with $J^2(h) \equiv \int_0^1 \left(h^{(m)}(w)\right)^2 dw$ and the integer $1 \leq m < \infty$ being known, and where $\epsilon_i \equiv (\epsilon_{i1}, \ldots \epsilon_{im})'$ is mean zero Gaussian with nonsingular and finite covariance $\Sigma_0$. The space $\mathcal{H}$ and function $J$ were defined previously in Sections 4.5 and 15.1 for the partly linear logistic regression model.

Let $\hat{V}_n$ be a symmetric, positive semidefinite $q \times q$ matrix estimator that converges in distribution to a symmetric, finite and positive definite matrix $V_0$. We will study estimation of $\beta_0$ and $h_0$ via maximizing the following objective function:

(22.14)

$$(\beta, h) \mapsto n^{-1} \sum_{i=1}^{n} (Y_i - Z_i\beta - h(W_i))' \hat{V}_n (Y_i - Z_i\beta - h(W_i)) + \lambda_n^2 J^2(h),$$

where $h(W_i) \equiv (h(W_{i1}), \ldots, h(W_{iq}))'$ and $\lambda_n > 0$ is a tuning parameter satisfying $\lambda_n = O_P(n^{-m/(2m+1)})$ and $\lambda_n^{-1} = O_P(n^{m/(2m+1)})$. Accordingly, let $\hat{\theta}_n \equiv (\hat{\beta}_n, \hat{h}_n)$ be a maximizer of (22.14).

We will establish consistency and rates of convergence for $\hat{\theta}_n$ as well as asymptotic normality and efficiency for $\hat{\beta}_n$. Note that the issue of asymptotic normality and efficient estimation of $\beta_0$ is nontrivial in this repeated measures context and that significant difficulties can arise. For instance, asymptotic normality may be difficult or impossible to achieve with apparently reasonable objective functions when $\hat{V}_n$ is not diagonal (see Lin and Carroll, 2001). A general efficient procedure for a variety of partly linear models for repeated measures, that includes our example as a special case, is given in Lin and Carroll (2005). However, the presentation and theory for this general approach, which the authors refer to as "profile kernel and backfitting estimation" is lengthy, and we will not include it here.

We now return to establishing the theoretical properties of $\hat{\theta}_n$. We need some additional assumptions before continuing. We assume common marginal expectations

$$\mathrm{E}\left[Z_{i1k} | W_{i1} = w\right] = \cdots = \mathrm{E}\left[Z_{iqk} | W_{iq} = w\right] \equiv \tilde{h}_k \in \mathcal{H},$$

for $k = 1, \ldots, p$. When $\hat{V}_n$ is not diagonal, we also require that $\mathrm{E}\left[Z_{ij} | W_i\right] = \mathrm{E}\left[Z_{ij} | W_{ij}\right]$, $j = 1, \ldots, q$. The first assumption is satisfied if the covariates $(Z_{ij}, W_{ij})$ have the same marginal distribution across $j = 1, \ldots, q$ and the

conditional expectation functions $(\tilde{h}_1, \ldots, \tilde{h}_p)$ are sufficiently smooth. The second assumption is satisfied if the components of $W_i$ other than $W_{ij}$ do not carry additional information for $Z_{ij}$, $j = 1, \ldots, q$.

Let $\zeta_{ij} \equiv (Z_{ij1}, \ldots, Z_{ijp}, 1, W_{ij}, W_{ij}^2, \ldots, W_{ij}^{m-1})'$. We need to assume

$$\begin{aligned} &P[\zeta_{ij}\zeta_{ij}'] \text{ is positive definite, and} \\ &P\|Z_{ij}\|^2 < \infty \text{ all } j = 1, \ldots, q. \end{aligned} \tag{22.15}$$

Note that this assumption precludes including an intercept in $Z_{ij}$. The intercept term for our model is contained in the function $h_0$. Define also

$$\tilde{Z}_{ij} \equiv Z_{ij} - \begin{pmatrix} \tilde{h}_1(W_i) \\ \vdots \\ \tilde{h}_p(W_i) \end{pmatrix}.$$

Our final assumptions are that $P[\tilde{Z}_{ij}' V_0 \tilde{Z}_{ij}]$ is nonsingular and that the density of $W_{ij}$ is bounded below by zero for at least one value of $j = 1, \ldots, q$. The assumptions in this paragraph are essentially boundedness and identifiability conditions on the covariates that will be needed for rate determination of both $\hat{\theta}_n$ and $\hat{h}_n$ and asymptotic normality. Unlike with previous examples in this book that involve partly linear models, we do not require any specific knowledge of the bounded sets containing $\beta_0$ and $h_0$ (compare, for example, with the first paragraph of Section 15.1). Neither do we require the coviarate vectors $Z_{ij}$ to be uniformly bounded, only bounded in an $L_2(P)$ sense. These weaker assumptions are more realistic in practice. However, more refined empirical process analysis is required.

*Consistency and rate calculation*

It turns out that in this example it is easier to compute the rate directly. Consistency will then follow. Define $U_{ij} \mapsto g_n(U_{ij}) \equiv \hat{\beta}_n' Z_{ij} + \hat{h}_n(W_{ij})$, $U_{ij} \mapsto g_0(U_{ij}) \equiv \beta_0' Z_{ij} + h_0(W_{ij})$,

$$\|g_n - g_0\|_{n,j} \equiv \left( n^{-1} \sum_{i=1}^{n} |(g_n - g_0)(U_{ij})|^2 \right)^{1/2},$$

and $\|g_n - g_0\|_n \equiv \left( \sum_{j=1}^{q} \|g_n - g_0\|_{n,j}^2 \right)^{1/2}$. Note that by definition of $\hat{\theta}_n$, we have

$$\begin{aligned} n^{-1} \sum_{i=1}^{n} \left( \epsilon_i - \begin{Bmatrix} (g_n - g_0)(U_{i1}) \\ \vdots \\ (g_n - g_0)(U_{iq}) \end{Bmatrix} \right)' \hat{V}_n \left( \epsilon_i - \begin{Bmatrix} (g_n - g_0)(U_{i1}) \\ \vdots \\ (g_n - g_0)(U_{iq}) \end{Bmatrix} \right) \\ + \lambda_n^2 J^2(\hat{h}_n) \;\leq\; n^{-1} \sum_{i=1}^{n} \epsilon_i' \hat{V}_n \epsilon_i + \lambda_n^2 J^2(h_0). \end{aligned}$$

Hence

$$(22.16)\quad n^{-1}\sum_{i=1}^{n}\begin{bmatrix}(g_n-g_0)(U_{i1})\\ \vdots\\ (g_n-g_0)(U_{iq})\end{bmatrix}'\hat{V}_n\begin{bmatrix}(g_n-g_0)(U_{i1})\\ \vdots\\ (g_n-g_0)(U_{iq})\end{bmatrix}+\lambda_n^2J^2(\hat{h}_n)$$
$$\leq\quad 2n^{-1}\sum_{i=1}^{n}\epsilon_i'\hat{V}_n\begin{bmatrix}(g_n-g_0)(U_{i1})\\ \vdots\\ (g_n-g_0)(U_{iq})\end{bmatrix}+\lambda_n^2J^2(h_0).$$

Combining (22.16) with our assumptions on $\hat{V}_n$ which ensure that $\hat{V}_n$ is asymptotically bounded and that its minimum eigenvalue is asymptotically bounded away from zero, we obtain

(22.17)

$$\|g_n-g_0\|_n^2+\lambda_n^2J^2(\hat{h}_n)\leq O_P\left(\lambda_n^2+\sum_{j=1}^{q}\sum_{l=1}^{q}n^{-1}\sum_{i=1}^{n}\epsilon_{il}(g_n-g_0)(U_{ij})\right).$$

The Cauchy-Schwartz inequality applied to the right-hand-side yields that

$$\|g_n-g_0\|_n^2\leq O_P(1+\|g_n-g_0\|_n),$$

and thus $\|g_n-g_0\|_n=O_P(1)$.

We will now proceed to the next step in which we utilize careful empirical process calculations to obtain tighter bounds for each of the $q^2$ terms of the form $n^{-1}\sum_{i=1}^{n}\epsilon_{il}(g_n-g_0)(U_{ij})$, where $1\leq j,l\leq q$. Accordingly, fix $j$ and $l$, and note that $\epsilon_{il}$ is mean zero Gaussian and independent of $U_{ij}$, $i=1,\ldots,n$. We will need to use the following theorem, which is Lemma 8.4 of van de Geer (2000), the proof of which we will omit:

THEOREM 22.7 *Let $v_1,\ldots,v_n$ be points on a space $\mathcal{V}$, and let $\mathcal{F}$ be a class of measurable functions on $\mathcal{V}$ with $\sup_{f\in\mathcal{F}}\|f\|_{Q_n}\leq R<\infty$, where $\|f\|_{Q_n}\equiv\left(n^{-1}\sum_{i=1}^{n}f^2(v_i)\right)^{1/2}$, and with*

$$\log N(\delta,\mathcal{F},Q_n)\leq A\delta^{-\alpha},\ \textit{all}\ \delta>0,$$

*for constants $0<\alpha<2$ and $A<\infty$. Suppose also that $\eta_1,\ldots,\eta_n$ are independent, mean zero real random variables satisfying*

$$\max_{1\leq i\leq n}K^2\left(\mathrm{E}e^{\eta_i^2/K^2}-1\right)\leq\sigma_0^2,$$

*for constants $0<K,\sigma_0<\infty$. Then for a constant $c$ which depends only on $A,\alpha,R,K,\sigma_0$ (and not on the points $v_1,\ldots,v_n$), we have for all $T>c$,*

$$\mathrm{P}\left(\sup_{f\in\mathcal{F}}\frac{\left|n^{-1/2}\sum_{i=1}^{n}\eta_if(v_i)\right|}{\|f\|_{Q_n}^{1-\alpha/2}}\geq T\right)\leq c\exp[-T^2/c^2].$$

Fix the sample $U_{1j} = u_{1j}, \ldots, U_{nj} = u_{nj}$, and set $v_i = u_{ij}$, $i = 1, \ldots, n$. Define $u \mapsto g_\theta(u) \equiv (\beta - \beta_0)'z + (h - h_0)(w)$, where $u \equiv (z', w)'$, $z \in \mathbb{R}^p$ and $w \in [0, 1]$. Consider the class of functions

$$\mathcal{F} \equiv \left\{ \frac{g_\theta(u)}{1 + \|g_\theta\|_{Q_n} + J(h - h_0)} : \theta \in \mathbb{R}^p \times \mathcal{H} \right\}.$$

By Taylor expansion,

$$\begin{aligned}
(h - h_0)(w) &= (h - h_0)(0) + (h - h_0)^{(1)}(0)w \\
&\quad + \ldots + \frac{(h - h_0)^{(m-1)} w^{m-1}}{(m-1)!} \\
&\quad + \int_0^w \frac{(h - h_0)^{(m)}(t)(w - t)^{m-1}}{(m-1)!} dt \\
&\equiv \sum_{k=0}^{m-1} h_k^* w^k + h_2(w) \\
&\equiv h_1(w) + h_2(w),
\end{aligned}$$

where $h_1$ is an $m-1$-degree polynomial and $\sup_{w \in [0,1]} |h_2(w)| \leq J(h - h_0)$. Thus

$$\|(\beta - \beta_0)'z + h_1(w)\|_{Q_n} \leq \|g_\theta\|_{Q_n} + J(h - h_0),$$

which implies

$$1 + \|g_\theta\|_{Q_n} + J(h - h_0) \geq 1 + \gamma_n \sqrt{\|\beta - \beta_0\|^2 + \sum_{k=0}^{m-1} (h_k^*)^2},$$

where $\gamma_n$ is the smallest eigenvalue of $B_n \equiv n^{-1} \sum_{i=1}^n \tilde{\zeta}_n \tilde{\zeta}_i'$, where $\tilde{\zeta}_i \equiv (z_{ij1}, \ldots, z_{ijp}, 1, w_{ij}, \ldots, w_{ij}^{m-1})'$, $i = 1, \ldots, n$. Using related arguments, it is also easy to verify that

$$\sup_{w \in [0,1]} |(h - h_0)(u)| \leq \left( \sum_{k=0}^{m-1} (h_k^*)^2 \right)^{1/2} + J(h - h_0).$$

Thus, for the class of functions

$$\mathcal{F}_1 \equiv \left\{ \frac{(h - h_0)(w)}{1 + \|g_\theta\|_{Q_n} + J(h - h_0)} : \beta \in \mathbb{R}^p, h \in \mathcal{H} \right\},$$

we have

$$\sup_{f \in \mathcal{F}_1, w \in [0,1]} |f(w)| \leq \gamma_n^{-1} + 1, \tag{22.18}$$

and, trivially,

$$\sup_{f \in \mathcal{F}_1} J(f) \leq 1. \tag{22.19}$$

We take a brief digression now to verify that without loss of generality, we can replace $\gamma_n^{-1}$ in (22.18) with a finite constant that does not depend on the particular values of $U_{1j}, \ldots, U_{nj}$. Assumption (22.15) ensures that $C_n \equiv n^{-1} \sum_{i=1}^n \zeta_{ij} \zeta_{ij}'$ converges almost surely to a positive definite matrix. Thus the minimum eigenvalue $\hat{\gamma}_n$ of $C_n$ will also converge to a positive constant. Thus, with probability going to 1 as $n \to \infty$, all possible values of $\gamma_n^{-1}$ will be bounded above by some $\Gamma_0 < \infty$. Since the $L_2(Q)$ norm is bounded by the uniform norm for any choice of probability measure $Q$, Theorem 9.21 now implies that

$$\log N(\delta, \mathcal{F}_1, Q_n) \leq A_0 \delta^{-1/m}, \text{ for all } \delta > 0, \tag{22.20}$$

where $A_0$ does not depend on the values of $u_{1j}, \ldots, u_{nj}$ (with probability tending to 1 as $n \to \infty$).

Now consider the class of functions

$$\mathcal{F}_2 \equiv \left\{ \frac{(\beta - \beta_0)' z}{1 + \|g_\theta\|_{Q_n} + J(h - h_0)} : \beta \in \mathbb{R}^p, h \in \mathcal{H} \right\},$$

and note that by previous arguments,

$$\mathcal{F}_2 \subset \mathcal{F}_3 \equiv \left\{ \beta' z : \beta \in \mathbb{R}^p, \|\beta\| \leq \gamma_n^{-1} \right\}.$$

Combining the fact that $n^{-1} \sum_{i=1}^n \|Z_{ij}\|^2$ converges almost surely to a finite constant by assumption with the previously established properties of $\gamma_n$, we obtain that

$$N(\delta, \mathcal{F}_2, Q_n) \leq A_1 \delta^{-p}, \text{ for all } \delta > 0, \tag{22.21}$$

where $A_1$ does not depend on the values of $u_{1j}, \ldots, u_{nj}$ (with probability tending to 1 as $n \to \infty$).

Combining (22.20) and (22.21), we obtain that

$$\log N(\delta, \mathcal{F}, Q_n) \leq A_2 \delta^{-1/m}, \text{ for all } \delta > 0,$$

where $A_2 < \infty$ does not depend on the values of $u_{1j}, \ldots, u_{nj}$ with probability tending to 1 as $n \to \infty$. Now we return to arguing for fixed $u_{1j}, \ldots, u_{nj}$ and apply Theorem 22.7 for $\alpha = 1/m$ to obtain that

$$\begin{aligned} &\mathrm{P}\left( \sup_{f \in \mathcal{F}} \frac{\left| n^{-1/2} \sum_{i=1}^n \epsilon_{il} f(u_{ij}) \right|}{\|f\|_{Q_n}^{1 - 1/(2m)}} \geq T \middle| U_{1j} = u_{1j}, \ldots, U_{nj} = u_{nj} \right) \\ &\qquad \leq \quad c \exp[-T^2/c^2], \end{aligned} \tag{22.22}$$

where $T$ and $c$ can be chosen to be both finite and not dependent on the values of $u_{1j}, \ldots, u_{nj}$. Accordingly, we can remove the conditioning in (22.22) and apply the fact that $\|g_n - g_0\|_{n,j} = O_P(1)$ to obtain

(22.23)
$$\left| n^{-1} \sum_{i=1}^{n} \epsilon_{il}(g_n - g_0)(u_{ij}) \right| = O_P\left( n^{-1/2} \|g_n - g_0\|_{n,j}^{1-1/(2m)} (1 + J(\hat{h}_n - h_0))^{1/(2m)} \right)$$

(see Exercise 14.6.9).

We now combine (22.23) with (22.17) to obtain that

(22.24)
$$\|g_n - g_0\|_n^2 + \lambda_n^2 J^2(\hat{h}_n)| = O_P\left( \lambda_n^2 + n^{-1/2} \|g_n - g_0\|_n^{1-1/(2m)} (1 + J(\hat{h}_n))^{1/(2m)} \right).$$

Putting $\tilde{R}_n \equiv n^{m/(2m+1)} \|g_n - g_0\|_n / (1 + J(\hat{h}_n))$, (22.24) now implies that $\tilde{R}_n^2 = O_P(1 + \tilde{R}_n^{1-1/(2m)})$. This implies that $\tilde{R}_n = O_P(1)$, and thus $\|g_n - g_0\|_n = O_P(n^{-m/(2m+1)}(1 + J(\hat{h}_n)))$. Apply this to (22.24), we obtain $J^2(\hat{h}_n) = O_P(1 + J(\hat{h}_n))$, which implies $J(\hat{h}_n) = O_P(1)$ and hence also $\|g_n - g_0\|_n = O_P(n^{-m/(2m+1)})$.

Recycling the arguments accompanying the polynomial expansion mentioned previously for $h \in \mathcal{H}$, we can verify that

$$\|\hat{\beta}_n - \beta_0\| = O_P(1) \quad \text{and} \quad \sup_{w \in [0,1]} |\hat{h}_n(w)| = O_P(1) \tag{22.25}$$

(see Exercise 22.6.10). Using this and recycling again previous arguments, we now have (see again Exercise 22.6.10) that $g_n - g_0$ is contained in a class $\mathcal{G}$ with probability increasing to 1, as $n \to \infty$, that satisfies

$$\log N_{[]}(\delta, \mathcal{G}, L_2(P)) \leq \tilde{A} \delta^{-1/m}, \text{ for all } \delta > 0. \tag{22.26}$$

We now use (22.26) to transfer the convergence rate for the norm $\|\cdot\|_{n,j}$ to the norm $\|\cdot\|_{P,2}$. This transfer follows, after setting $\nu = 1/m$, from the following theorem, which is Theorem 2.3 of Mammen and van de Geer (1997) and which we present without proof:

THEOREM 22.8 *Suppose the class of measurable functions $\mathcal{F}$ satisfies the following for some norm $\|\cdot\|$, constants $A < \infty$ and $0 < \nu < 2$, and for all $\delta > 0$:*

$$\log N_{[]}(\delta, \mathcal{F}, \|\cdot\|) \leq A \delta^{-\nu}.$$

*Then for all $\eta > 0$ there exists a $0 < C < \infty$ such that*

$$\limsup_{n \to \infty} \mathrm{P}\left( \sup_{f \in \mathcal{F}, \|f\| > C n^{-1/(2+\nu)}} \left| \frac{\|f\|_n}{\|f\|} - 1 \right| > \eta \right) = 0.$$

It now follows (see Exercise 22.6.12) that

$$\|g_n - g_0\|_{P,2} = O_P(n^{-m/(2m+1)}). \tag{22.27}$$

Now (22.27) implies for each $j = 1, \ldots, q$ that

$$P\left((\hat{\beta}_n - \beta_0)' Z_{ij} + (\hat{h}_n - h_0)(W_{ij})\right)^2 = O_P(n^{-2m/(2m+1)}),$$

which, by the assumptions and definition of $\tilde{Z}_{ij}$, implies for some function $f_n$ of $W_{ij}$ that

$$\begin{aligned} O_P(n^{-2m/(2m+1)}) &= P\left((\hat{\beta}_n - \beta_0)' \tilde{Z}_{ij} + \tilde{f}_n(W_{ij})\right)^2 \\ &\geq P\left((\hat{\beta}_n - \beta_0)' \tilde{Z}_{ij}\right)^2 \\ &\geq \tilde{c}\|\hat{\beta}_n - \beta_0\|^2, \end{aligned}$$

for some $\tilde{c} > 0$. Thus $\|\hat{\beta}_n - \beta_0\| = O_P(n^{-m/(2m+1)})$ and therefore also $P\left((\hat{h}_n - h_0)(W_{ij})\right)^2 = O_P(n^{-2m/(2m+1)})$. Since the density of $W_{ij}$ is bounded below by 0 for at least one value of $j \in \{1, \ldots, q\}$, we also have $\|\hat{h}_n - h_0\|_{L_2} = O_P(n^{-m/(2m+1)})$. The fact that $J(\hat{h}_n) = O_P(1)$ now, finally, yields $\|\hat{h}_n - h_0\|_\infty = o_P(1)$. Thus we have achieved uniform consistency of both parameters as well as the optimal $L_2$ rate of convergence for $\hat{h}_n$.

*Asymptotic normality*

Consider evaluating the objective function in (22.14) at the point $\theta_{n,s} \equiv \left(\hat{\beta}_n + st, \hat{h}_n - st'(\tilde{h}_1, \ldots, \tilde{h}_p)'\right)$, where $s$ is a scalar and $t \in \mathbb{R}^p$. Differentiating with respect to $s$, evaluating at $s = 0$ and allowing $t$ to range over $t \in \mathbb{R}^p$, we obtain by definition of a maximizer that

$$\begin{aligned} 0 &= n^{-1} \sum_{i=1}^n \left[\tilde{Z}_i - \tilde{H}(W_i)\right]' \hat{V}_n \left[Y_i - Z_i \hat{\beta}_n - \hat{h}_n(W_i)\right] \\ &\quad -\lambda_n^2 \int_0^1 \begin{bmatrix} \hat{h}_n(w)\tilde{h}_1(w) \\ \vdots \\ \hat{h}_n(w)\tilde{h}_p(w) \end{bmatrix} dw, \end{aligned}$$

where $\tilde{H}(W_i) \equiv (\tilde{h}_1(W_i), \ldots, \tilde{h}_p(W_i))$ and, for any function $g : [0,1] \mapsto \mathbb{R}$, $g(W_i) \equiv (g(W_{i1}), \ldots, g(W_{iq}))'$. This now implies

$$
\begin{aligned}
& n^{-1}\sum_{i=1}^{n}\left[\tilde{Z}_i-\tilde{H}(W-i)\right]'\hat{V}_n Z_i(\hat{\beta}_n-\beta_0) \\
&\quad = \quad n^{-1}\sum_{i=1}^{n}\left[\tilde{Z}_i-\tilde{H}(W_i)\right]'\hat{V}_n\epsilon_i \\
&\qquad -n^{-1}\sum_{i=1}^{n}\left[\tilde{Z}_i-\tilde{H}(W_i)\right]'\hat{V}_n\left[(\hat{h}_n-h_0)(W_i)\right] \\
&\qquad -\lambda_n^2\int_0^1\begin{bmatrix}\hat{h}_n(w)\tilde{h}_1(w)\\ \vdots \\ \hat{h}_n(w)\tilde{h}_p(w)\end{bmatrix}dw \\
&\quad \equiv \quad A_n-B_n-C_n.
\end{aligned}
$$

We save it as an exercise (see Exercise 22.6.13) to verify that

$$
\begin{aligned}
n^{-1}\sum_{i=1}^{n}\left[\tilde{Z}_i-\tilde{H}(W_i)\right]'\hat{V}_n Z_i \quad &\xrightarrow{\mathrm{P}} \quad P\left[\tilde{Z}_i'V_0\tilde{Z}_i\right], \\
\sqrt{n}A_n \quad &\rightsquigarrow \quad N_p\left(0, P\left[\tilde{Z}_i'V_0\Sigma_0V_0\tilde{Z}_i\right]\right), \\
B_n \;=\; o_P(n^{-1/2}), \quad &\text{and} \quad C_n \;=\; o_P(n^{-1/2}),
\end{aligned}
$$

where $N_p(a,B)$ is a $p$-variate normal distribution with mean vector $a$ and variance matrix $B$. Thus we can conclude

(22.28)

$$
\sqrt{n}(\hat{\beta}_n-\beta_0)\rightsquigarrow N_p\left(0,\left\{P[\tilde{Z}_i'V_0\tilde{Z}_i]\right\}^{-1}P\left[\tilde{Z}_i'V_0\Sigma_0V_0\tilde{Z}_i\right]\left\{P[\tilde{Z}_i'V_0\tilde{Z}_i]\right\}^{-1}\right).
$$

*Efficiency*

Based on the Gaussian likelihood, it is easy to verify that the score for $\beta$ is $Z_i'\Sigma_0^{-1}\epsilon_i$ and the tangent set for $h$ is $\{h'(W_i)\Sigma_0^{-1}\epsilon_i : h\in\mathcal{H}\}$. We could choose to take the closure of the tangent set in the space of square-integrable functions of $W_i$ and to center the set so that all elements have mean zero. However, these adjustments are not needed to derive that the efficient score for $\beta$ is

$$
\tilde{\ell}_{\beta,h}(X_i) \quad \equiv \quad [\tilde{Z}_i-\tilde{H}(W_i)]'\Sigma_0^{-1}\epsilon_i. \tag{22.29}
$$

The verification of this is saved for Exercise 22.6.8.

Accordingly, choosing $\hat{V}_n$ so that it converges to $V_0=\Sigma_0^{-1}$ will result in optimal estimation of $\hat{V}_n$. This comes as no surprise given the form of the limiting variance of $\sqrt{n}(\hat{\beta}_n-\beta_0)$ in (22.28) and the optimality concepts of Godambe (1960) (see Proposition 4.3 of Section 4.4 and the surrounding discussion). The resulting limiting covariance of $\sqrt{n}(\hat{\beta}_n-\beta_0)$ is $\{P[\tilde{Z}_i'\Sigma_0^{-1}\tilde{Z}_i]\}^{-1}$. The following choice for $\hat{V}_n$ will work:

(22.30)

$$\hat{V}_n = \left[ n^{-1} \sum_{i=1}^{n} \left( Y_i - Z_i \check{\beta}_n - \check{h}_n(W_i) \right) \left( Y_i - Z_i \check{\beta}_n - \check{h}_n(W_i) \right)' \right]^{-1},$$

where $\check{\theta}_n \equiv (\check{\beta}_n, \check{h}_n)$ is a minimizer of (22.14) with $\hat{V}_n$ replaced by the $q \times q$ identity matrix. Verification that $\hat{V}_n \xrightarrow{\mathrm{P}} \Sigma_0^{-1}$ is saved for Exercise 22.6.14.

Efficient estimation of $\hat{\beta}_n$ using $\hat{V}_n$ given in (22.30) thus requires two optimizations of an objective function, but this is a small price to pay for efficiency. Note that more information on the structure of $\Sigma_0$ could be used to reduce the number of parameters needed for estimating $\hat{V}_n$. For example, identity, diagonal, exchangeable or autocorrelation covariance structures could be employed. If the covariance structure is correctly specified, full efficiency is obtained. If it is not correctly specified, asymptotic normality of $\sqrt{n}(\hat{\beta}_n - \beta_0)$ is achieved although the associated limiting variance is suboptimal. This is a very simple example of a "locally efficient estimator" as mentioned briefly at the end of Chapter 16 (see also Section 1.2.6 of van der Laan and Robins, 2003).

Moreover, it seems quite reasonable that efficient estimation of $\Sigma_0$ could also be developed, but we do not pursue this here. We also have not yet discussed a method of estimating the limiting covariance of $\sqrt{n}(\hat{\beta}_n - \beta_0)$. However, inference based on weighted M-estimation under penalization, as discussed in Section 21.5, appears to be applicable here. We omit the details.

One final issue we mention is in regards to the assumption we make that $\mathrm{E}[Z_{ij}|W_i] = \mathrm{E}[Z_{ij}|W_{ij}]$, $j = 1, \ldots, q$, $i = 1, \ldots, n$. This seems oddly restrictive, but it definitely makes the asymptotic normality result easier to achieve when $\hat{V}_n$ is not diagonal. The complications resulting from omitting this assumption are connected to several intriguing observations made in Lin and Carroll (2001). The profile kernel and backfitting estimation method of Lin and Carroll (2006) seems to be able to get around this problem so that full efficiency is achieved under less restrictive assumptions on the dependency between $Z_{ij}$ and $W_{ij}$ than what we have used here. We refer the interested reader to the cited publications for more details.

## 22.5 Proofs

**Proof of Theorem 22.1.** We will use Theorem 9.20. Let $\gamma = \alpha \vee \beta + 3/2$, and note that

$$\begin{aligned} \frac{|f(x)|}{1+|x|^{\gamma}} &\leq \frac{|f(x)|}{1+|x|^{\alpha}} \times 2 \frac{1+|x|^{\alpha}}{1+|x|^{\alpha+3/2}} \\ &\leq M \times 6(1 \vee |x|)^{-3/2}, \end{aligned}$$

for all $x \in \mathbb{R}$. Similarly, we can verify that

$$\frac{|\dot{f}(x)|}{1+|x|^\gamma} \leq 6M(1 \vee |x|)^{-3/2},$$

for all $x \in \mathbb{R}$.

Now define the class of functions

$$\mathcal{G} \equiv \left\{ x \mapsto \frac{f(x)}{1+|x|^\gamma} : f \in \mathcal{F} \right\},$$

and note that for every $g \in \mathcal{G}$ and all $x \in \mathbb{R}$, $|g(x)| \leq 6M(1 \vee |x|)^{-3/2}$ and

$$\begin{aligned} |\dot{g}(x)| &\leq \frac{|\dot{f}(x)|}{1+|x|^\gamma} + \frac{|f(x)|(1 \vee |x|)^{\gamma-1}}{(1+|x|^\gamma)^2} \\ &\leq 6M(1 \vee |x|)^{-3/2} + 6M(1 \vee |x|)^{-3/2}. \end{aligned}$$

Setting $k_* \equiv 12M$, we obtain that $|g(x)| \vee |\dot{g}(x)| \leq k_*(1 \vee |x|)^{-3/2}$.

Now let $I_{j,+} \equiv [j, j+1]$ and $I_{j,-} \equiv [-(j+1), -j]$ for all integers $j \geq 0$, and note that each of these sets is trivially bounded and convex and that their union equals $\mathbb{R}$. Let $\|\cdot\|_1$ be the Lipschitz norm of degree 1, and note that for every $g \in \mathcal{G}$, $\left\|g_{|I_{j,+}}\right\|_1 \leq k(1 \vee j)^{-3/2}$ for all $j \geq 0$, where $g_{|A}$ is the restriction of $g$ to the set $A \subset \mathbb{R}$. Similarly, $\left\|g_{|I_{j,-}}\right\|_1 \leq k_*(1 \vee j)^{-3/2}$. Note also that $I^1_{j,+} = \{x : \|x - I_{j,+}\| < 1\} = (j-1, j+2)$ and, similarly, $I^1_{j,-} = (-j-2, -j+1)$, for all $j \geq 0$.

Applying Theorem 9.20 and using the fact that the probability of any set is $\leq 1$, we obtain

$$\log N_{[]}(\delta, \mathcal{G}, L_4(Q)) \leq \left(\frac{K_*}{\delta}\right) \left(2(3k)^{4/5} \sum_{j=0}^{\infty} (1 \vee j)^{-6/5}\right)^{5/4},$$

for all $\delta > 0$, where $K_* < \infty$ is universal. Since $\sum_{j=1}^{\infty} j^{-6/5}$ converges, we obtain that there exists a constant $K_{**} < \infty$, depending only on $k_*$, such that

$$\log N_{[]}(\delta, \mathcal{G}, L_4(Q)) \leq \frac{K_{**}}{\delta}, \tag{22.31}$$

for all $\delta > 0$ and all probability measures $Q$.

Note that for any $f_1, f_2 \in \mathcal{F}$,

$$\begin{aligned} Q[f_1 - f_2]^2 &= Q[(1+|X|^\gamma)^2 (g_1 - g_2)^2] \\ &\leq \left(Q[1+|X|^\gamma]^4]\right)^{1/2} \left(Q[g_1 - g_2]^4\right)^{1/2}, \end{aligned}$$

where $x \mapsto g_j(x) \equiv (1+|x|^\gamma)^{-1} f_j(x)$, for $j = 1, 2$. Suppose $Q|X|^{4\gamma} = P|X|^{4(\alpha \vee \beta)+6} \equiv M_* < \infty$. Then $Q(1+|X|^\gamma)^4 \leq 8(1+M_*)$. Now let

$[\{g_{1i}, g_{2i}\}, 1 \le i \le m]$ be a minimal covering of $\delta$-$L_4(Q)$ brackets for $\mathcal{G}$. Then $[\{(1+|x|^\gamma)g_{1i}, (1+|x|^\gamma)g_{2i}\}, 1 \le i \le m]$ is an $[8(1+M_*)]^{1/4}\delta$-$L_2(Q)$ covering of $\mathcal{F}$. The desired conclusion of the Theorem now follows by (22.31) and the fact that $k_*$ depends only on $\alpha, \beta, M$.□

**Proof of Corollary 22.2.** It is not hard to verify (see Exercise 22.6.14) that the class

$$\mathcal{H}_0 \equiv \{-(\beta-\beta_0)'Z : \beta \in B\} \tag{22.32}$$

satisfies

$$N_{[]}(\delta, \mathcal{H}_0, L_2(P)) \le k_{**}\delta^{-k}, \tag{22.33}$$

for all $\delta > 0$ and some constant $k_{**} < \infty$ (which may depend on $P$).

Fix $P$ that satisfies the conditions of the corollary, and let $\epsilon \equiv Y - \beta_0'Z$. For any $h \in \mathcal{H}_0$, define $Y_h \equiv \epsilon + h(Z)$ and let $Q_h$ be the probability distribution for $Y_h$. Note that $\sup_{h\in\mathcal{H}_0} Q_h|Y_h|^{4(\alpha\vee\beta)+6} \equiv M_{**} < \infty$. Now fix $\delta > 0$, and let $\{h_{1i}, h_{2i}\}$, $i = 1, \ldots, m$ be a minimal collection of $\delta/3$-$L_2(P)$ brackets that cover $\mathcal{H}_0$. Let $H_\delta$ be the collection of all of the functions $h_{ji}$, $j = 1, 2$, $i = 1, \ldots, m$, and let $h \in H_\delta$ be arbitrary. Theorem 22.1 now implies that

$$\log N_{[]}(\delta/3, \mathcal{F}, L_2(Q_h)) \le \frac{K_0}{\delta},$$

for $K_0 < \infty$ that does not depend on $\delta$ nor on $Q_h$. Accordingly, let $\{f_{1i,h}, f_{2i,h}\}$, $i = 1, \ldots, m_*$ be a minimal collection of $\delta/3$-$L_2(Q_h)$ brackets that covers $\mathcal{F}$. Moreover, by definition of $Q_h$, $\left(P[f_{2i,h}(Y) - f_{1i,h}]^2\right)^{1/2} \le \delta/3$. Repeat this process for all members of $H_\delta$.

Let $f(Y - \beta'Z)$ be an arbitrary element of $\mathcal{H}$, and let $\{h_{1j}, h_{2j}\}$ be the $\delta/3$-$L_2(P)$ bracket that contains $-(\beta-\beta_0)'Z$. Let $\{f_{1i,h_{1j}}, f_{2i,h_{1j}}\}$ and $\{f_{1i',h_{2j}}, f_{2i',h_{2j}}\}$ be $\delta/3$-$L_2(Q_{h_{1j}})$ and $\delta/3$-$L_2(Q_{h_{2j}})$ brackets, respectively, that cover $h_{1j}$ and $h_{2j}$. Then

$$f_{1i,h_{1j}}(Y + h_{1j}(Z)) \le f(Y - \beta'Z) \le f_{2i',h_{2j}}(Y + h_{2j}(Z))$$

and

$$\begin{aligned}
&\left(P\left[f_{2i',h_{2j}}(Y + h_{2j})(Z)) - f_{1i,h_{1j}}(Y + h_{1j}(Z))\right]^2\right)^{1/2} \\
&\quad \le \left(Q_{h_{2j}}\left[f_{1i',h_{2j}} - f_{2i',h_{2j}}\right]^2\right)^{1/2} + \delta/3 + \left(Q_{h_{1j}}\left[f_{1i,h_{1j}} - f_{2i,h_{1j}}\right]^2\right)^{1/2} \\
&\quad \le \delta.
\end{aligned}$$

Since the logarithm of the number of such brackets is bounded by $K_0'\delta^{-1}$, for a constant $K_0'$ depending on $P$ but not on $\delta$, and since $\delta$ was arbitrary, we have that the bracketing entropy integral for $\mathcal{H}$, $J_{[]}(\infty, \mathcal{H}, L_2(P))$, is finite. Thus the desired conclusion follows from Theorem 2.3.□

## 22.6 Exercises

**22.6.1.** Verify that $\mathcal{F}_1$ given in (22.3) is Donsker and that $\mathcal{F}_2$ given in (22.4) is Glivenko-Cantelli.

**22.6.2.** Show that the conditions given in the second paragraph for the residual density $\eta$ are satisfied by the standard normal density, with $a_1 = 1$, $a_2 = 2$, and $b_1 = b_2 = 1$.

**22.6.3.** For the linear regression example, show that having $P[ZZ']$ be positive definite combined with having $x \mapsto (\dot{\eta}_0/\eta_0)(x)$ be strictly monotone (along with the other assumptions specified for this model) implies that the map $\beta \mapsto P\tilde{\ell}_{\beta,\eta_0}$ is continuous with a unique zero at $\beta = \beta_0$.

**22.6.4.** In the linear regression example, verify that both $\|\hat{\eta}_n^{(1)} - \dot{\eta}_0\|_\infty = O_P(n^{-1/2}h_n^{-2} + h_n^2)$ and $\|\hat{\eta}_n^{(2)} - \ddot{\eta}_0\|_\infty = O_P(n^{-1/2}h_n^{-3} + h_n)$.

**22.6.5.** Show that the estimator $\check{F}_n$ defined at the end of Section 22.2 is uniformly efficient.

**22.6.6.** Show that the piggyback bootstrap $(\beta_n, \check{F}_n^\circ)$ defined at the end of Section 22.2 is valid.

**22.6.7.** Consider a cadlag, monotone increasing stochastic process $\{X(t) : t \in T\}$, and consider the class of evaluation functions $\mathcal{F} \equiv \{f_t : t \in T\}$, where $f_t(X) \equiv X(t)$ for all $t \in T$, and $T$ is a closed interval in $\mathbb{R}$. Show that $\mathcal{F}$ is PM.

**22.6.8.** Verify that (22.23) follows from (22.22). Part of this verification should include a discussion and resolution of any measurability issues.

**22.6.9.** Verify that both (22.25) and (22.26) hold. Hint: For the second inequality, consider the classes $\{(\beta - \beta_0)'Z_{ij} : \|\beta - \beta_0\| \le K_1\}$ and $\{(h - h_0)(W_{ij}) : \|h - h_0\|_\infty \le K_2, J(h - h_0) \le K_3\}$ separately. For the first class, the desired bracketing entropy bound is easy to verify. For the second class, it may be helpful to first establish the bracketing entropy bound for the uniform norm and then apply Lemma 9.22.

**22.6.10.** Verify (22.27). It may be helpful to first verify it for fixed $j \in \{1, \ldots, q\}$, since $U_{1j}, \ldots, U_{nj}$ forms an i.i.d. sample.

**22.6.11.** Verify (22.27).

**22.6.12.** Utilizing arguments applied in Section 22.4 as needed, verify (22.28).

**22.6.13.** Show the following:

1. $\tilde{\ell}_{\beta,h}$ given in (22.29) is the efficient score for $\beta$ as claimed.
2. $\hat{V}_n$ given in (22.30) is consistent for $\Sigma_0^{-1}$.

**22.6.14.** Verify that the class $\mathcal{H}_0$ defined in (22.32) satisfies the bracketing entropy bound (22.33) for a constant $k_* < \infty$ that may depend on $P$ (but only through the compact set containing $Z$ $P$-almost surely).

## 22.7 Notes

Theorems 22.3, 22.5, and 22.6 are, after at most a few minor modifications, Theorems A1, A2 and A3 of Fine, Yan and Kosorok (2004). Other connections to original sources have been acknowledged previously.

# References


Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

Andrews, D. W. K. (1991). An empirical process central limit theorem for dependent non-identically distributed random variables. *Journal of Multivariate Analysis*, 38:187–203.

Andrews, D. W. K. (2001). Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica*, 69:683–673.

Andrews, D. W. K. and Ploberger, W. (1994). Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica*, 62:1383–1414.

Arcones, M. A. and Giné, E. (1993). Limit theorems for U-processes. *Annals of Probability*, 21:1494–1542.

Arcones, M. A. and Wang, Y. (2006). Some new tests for normality based on U-processes. *Statistics and Probability Letters*, 76:69–82.

Arcones, M. A. and Yu, B. (1994). Central limit theorems for empirical and U-processes of stationary mixing sequences. *Journal of Theoretical Probability*, 7:47–71.

Barbe, P. and Bertail, P. (1995). *The Weighted Bootstrap*. Springer, New York.

Bassett, G., Jr. and Koenker, R. (1978). Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association*, 73:618–622.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300.

Bickel, P. J., Götze, F., and van Zwet, W. R. (1997). Resampling fewer than $n$ observations: Gains, losses, and remedies for losses. *Statistica Sinica*, 7:1–31.

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, New York.

Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Annals of Statistics*, 1:1071–1095.

Bilius, Y., Gu, M. G., and Ying, Z. (1997). Towards a general theory for Cox model with staggered entry. *Annals of Statistics*, 25:662–682.

Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

Billingsley, P. (1986a). *Probability and Measure*. Wiley, New York, second edition.

Billingsley, P. (1986b). *Probability and Measure*. Wiley, New York, third edition.

Bradley, R. C. (1986). Basic properties of strong mixing conditions. In Eberlein, E. and Taqqu, M. S., editors, *Dependence in Probability and Statistics: A Survey of Recent Results*, pages 165–192. Birkhäuser, Basel.

Bretagnolle, J. and Massart, P. (1989). Hungarian construction from the nonasymptotic viewpoint. *Annals of Probability*, 17:239–256.

Bühlmann, P. (1995). The blockwise bootstrap for general empirical processes of stationary sequences. *Stochastic Processes and Their Applications*, 58:247–265.

Cai, T. X. and Betensky, R. A. (2003). Hazard regression for interval-censored data with penalized splines. *Biometrics*, 59:570–579.

Cantelli, F. P. (1933). Sulla determinazione empirica delle leggi di probabilità. *Giornale dell'Istituto Italiano degli Attuari*, 4:421–424.

Chang, M. N. (1990). Weak convergence of a self-consistent estimator of the survival function with doubly censored data. *Annals of Statistics*, 18:391–404.

Cheng, G. and Kosorok, M. R. General frequentist properties of the posterior profile distribution. *Annals of Statistics*. To appear.

Cheng, G. and Kosorok, M. R. Higher order semiparametric frequentist inference with the profile sampler. *Annals of Statistics*. To appear.

Cheng, G. and Kosorok, M. R. (2007). The penalized profile sampler. http://arxiv.org/abs/math.ST/0701540.

Conway, J. B. (1990). *A Course in Functional Analysis*. Springer, New York, 2nd edition.

Cox, D. D. and O'Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *Annals of Statistics*, 18:1676–1695.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62:269–276.

Dedecker, J. and Louhichi, S. (2002). Maximal inequalities and empirical central limit theorems. In Dehling, H., Mikosch, T., and Sørensen, M., editors, *Empirical Process Techniques for Dependent Data*, pages 137–159. Birkhäuser, Boston.

Dehling, H., Mikosch, T., and Sørensen, M., editors (2002). *Empirical Process Techniques for Dependent Data*. Birkhäuser, Boston.

Dehling, H. and Philipp, W. (2002). Empirical process techniques for dependent data. In Dehling, H., Mikosch, T., and Sørensen, M., editors, *Empirical Process Techniques for Dependent Data*, pages 3–113. Birkhäuser, Boston.

Dehling, H. and Taqqu, M. S. (1989). The empirical process of some long-range dependent sequences with an application to U-statistics. *Annals of Statistics*, 17:1767–1783.

Di Bucchiano, A., Einmahl, J. H. J., and Mushkudiani, N. A. (2001). Smallest nonparametric tonerance regions. *Annals of Statistics*, 29:1320–1343.

Dixon, J. R., Kosorok, M. R., and Lee, B. L. (2005). Functional inference in semiparametric models using the piggyback bootstrap. *Annals of the Institute of Statistical Mathematics*, 57:255–277.

Donsker, M. D. (1952). Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *Annals of Mathematical Statistics*, 23:277–281.

Dudley, R. M. and Philipp, W. (1983). Invariance principles for sums of Banach space valued random elements and empirical processes. *Probability Theory and Related Fields*, 62:509–552.

Dugundji, J. (1951). An extension of Tietze's theorem. *Pacific Journal of Mathematics*, 1:353–367.

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.

Eberlein, E. and Taqqu, M. S., editors (1986). *Dependence in Probability and Statistics: A Survey of Recent Results*. Birkhäuser, Basel.

Efron, B. (1967). The two sample problem with censored data. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 4:831–855.

Fine, J. P., Yan, J., and Kosorok, M. R. (2004). Temporal process regression. *Biometrika*, 91:683–703.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

Fleming, T. R., Harrington, D. P., and O'Sullivan, M. (1987). Supremum versions of the logrank and generalized wilcoxon statistics. *Journal of the American Statistical Association*, 82:312–320.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society, Series B*, 64:499–517.

Ghosal, S., Sen, A., and van der Vaart, A. W. (2000). Testing monotonicity of regression. *Annals of Statistics*, 28:1054–1082.

Gilbert, P. B. (2000). Large sample theory of maximum likelihood estimates in semiparametric biased sampling models. *Annals of Statistics*, 28:151–194.

Giné, E. (1996). Lectures on some aspects of the bootstrap. Unpublished notes.

Giné, E. (1997). Decoupling and limit theorems for U-statistics and U-processes. In *Lectures on probability theory and statistics (Saint-Flour, 1996)*, pages 1–35. Springer, Berlin. Lecture Notes in Mathematics, 1665.

Glivenko, V. (1933). Sulla determinazione empirica della leggi di probabilità. *Giornale dell'Istituto Italiano Degli Attuari*, 4:92–99.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, 31:1208–1211.

Grenander, U. (1956). On the theory of mortality measurement, part ii. *Skandinavisk Aktuarietidskrift*, 39:125–153.

Groeneboom, P. (1989). 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 Verlag, Basel.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer, New York.

Gu, M. G. and Lai, T. L. (1991). Weak convergence of time-sequential censored rank statistics with applications to sequential testing in clinical trials. *Annals of Statistics*, 19:1403–1433.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chaines and their applications. *Biometrika*, 57:97–109.

Huang, J. (1996). Efficient estimation for the Cox model with interval censoring. *Annals of Statistics*, 24:540–568.

Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

Hunter, D. R. and Lange, K. (2002). Computing estimates in the proportional odds model. *Annals of the Institute of Statistical Mathematics*, 54:155–168.

Jameson, J. O. (1974). *Topology and Normed Spaces*. Chapman and Hall, London.

Johnson, W. B., Lindenstrauss, J., and Schechtman, G. (1986). Extensions of Lipschitz maps into Banach spaces. *Israel Journal of Mathematics*, 54:129–138.

Karatzas, I. and Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus*. Springer, New York, 2 edition.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *Annals of Mathematical Statistics*, 27:887–906.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Annals of Statistics*, 18:191–219.

Kim, Y. and Lee, J. (2003). Bayesian bootstrap for proportional hazards models. *Annals of Statistics*, 31:1905–1922.

Komlós, J., Major, P., and Tusnády, G. (1975). An approximation of partial sums of independent rv's and the sample df. i. *Probability Theory and Related Fields*, 32:111–131.

Komlós, J., Major, P., and Tusnády, G. (1976). An approximation of partial sums of independent rv's and the sample df. ii. *Probability Theory and Related Fields*, 34:33–58.

Kosorok, M. R. (1999). Two-sample quantile tests under general conditions. *Biometrika*, 86:909–921.

Kosorok, M. R. (2003). Bootstraps of sums of independent but not identically distributed stochastic processes. *Journal of Multivariate Analysis*, 84:299–318.

Kosorok, M. R., Lee, B. L., and Fine, J. P. (2004). Robust inference for univariate proportional hazards frailty regression models. *Annals of Statistics*, 32:1448–1491.

Kosorok, M. R. and Ma, S. (2005). Comment on "Semilinear high-dimensional model for normalization of microarray data: a theoretical analysis and partial consistency" by J. Fan, H. Peng, T. Huang. *Journal of the American Statistical Association*, 100:805–807.

Kosorok, M. R. and Ma, S. (2007). Marginal asymptotics for the "large p, small n" paradigm: with applications to microarray data. *Annals of Statistics*, 35:1456–1486.

Kosorok, M. R. and Song, R. (2007). Inference under right censoring for transformation models with a change-point based on a covariate threshold. *Annals of Statistics*, 35:957–989.

Kress, R. (1999). *Linear Integral Equations*. Springer, New York, second edition.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observation. *Annals of Statistics*, 17:1217–1241.

Lee, B. L. (2000). *Efficient Semiparametric Estimation Using Markov Chain Monte Carlo*. PhD thesis, University of Wisconsin at Madison, Madison, Wisconsin.

Lee, B. L., Kosorok, M. R., and Fine, J. P. (2005). The profile sampler. *Journal of the American Statistical Association*, 100:960–969.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22.

Lin, D. Y., Fleming, T. R., and Wei, L. J. (1994). Confidence bands for survival curves under the proportional hazards model. *Biometrika*, 81:73–81.

Lin, X. and Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association*, 96:1045–1056.

Lin, X. and Carroll, R. J. (2006). Semiparametric estimation in general repeated measures problems. *Journal of the Royal Statistical Society, Series B*, 68:69–88.

Liu, R. Y. and Singh, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In LePage, R. and Billard, L., editors, *Exploring the Limits of Bootstrap*, pages 225–248. Wiley, New York.

Ma, S. and Kosorok, M. R. (2005a). Penalized log-likelihood estimation for partly linear transformation models with current status data. *Annals of Statistics*, 33:2256–2290.

Ma, S. and Kosorok, M. R. (2005b). Robust semiparametric M-estimation and the weighted bootstrap. *Journal of Multivariate Analysis*, 96:190–217.

Ma, S., Kosorok, M. R., Huang, J., Xie, H., Manzella, L., and Soares, M. B. (2006). Robust semiparametric microarray normalization and significance analysis. *Biometrics*, 62:555–561.

Mammen, E. and van de Geer, S. A. (1997). Penalized quasi-likelihood estimation in partial linear models. *Annals of Statistics*, 25:1014–1035.

Mason, D. M. and Newton, M. A. (1992). A rank statistic approach to the consistency of a general boostrap. *Annals of Statistics*, 25:1611–1624.

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Annals of Probability*, 18:1269–1283.

Megginson, R. E. (1998). *An Introduction to Banach Space Theory*. Springer, New York.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092.

Murphy, S. A. (1994). Consistency in a proportional hazards model incorporating a random effect. *Annals of Statistics*, 22:712–731.

Murphy, S. A., Rossini, A. J., and van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, 92:968–976.

Murphy, S. A. and van der Vaart, A. W. (1999). Observed information in semiparametric models. *Bernoulli*, 5:381–412.

Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. with comments and a rejoinder by the authors. *Journal of the American Statistical Association*, 95:449–485.

Nadeau, C. and Lawless, J. F. (1998). Inferences for means and covariances of point processes through estimating functions. *Biometrika*, 85:893–906.

Naik-Nimbalkar, U. V. and Rajarshi, M. B. (1994). Validity of blockwise bootstrap for empirical processes with stationary observations. *Annals of Statistics*, 22:980–994.

Neumeyer, N. (2004). A central limit theorem for two-sample U-processes. *Statistics and Probability Letters*, 67:73–85.

Nolan, D. and Pollard, D. (1987). U-processes: rates of convergence. *Annals of Statistics*, 15:780–799.

Nolan, D. and Pollard, D. (1988). Functional limit theorems for U-processes. *Annals of Probability*, 16:1291–1298.

Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *Annals of Statistics*, 26:183–214.

Peligrad, M. (1998). On the blockwise bootstrap for empirical processes for stationary sequences. *Annals of Probability*, 26:877–901.

Pfanzagl, J. (1990). *Estimation in Semiparametric Models: Some Recent Developments*, volume 63 of *Lecture Notes in Statistics*. Springer.

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsampling under minimal assumptions. *Annals of Statistics*, 22:2031–2050.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*, volume 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics and American Statistical Association, Hayward, California.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7:186–199.

Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters-An excess mass approach. *Annals of Statistics*, 23:855–881.

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.

Prentice, R. L. and Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, 47:825–839.

Radulović, D. (1996). The bootstrap for empirical processes based on stationary observations. *Stochastic Processes and Their Applications*, 65:259–279.

Rubin, D. B. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9:130–134.

Rudin, W. (1987). *Real and Complex Analysis*. McGraw-Hill, Inc., New York, third edition.

Rudin, W. (1991). *Functional Analysis*. McGraw-Hill, Inc., New York, second edition.

Scheffé, H. (1959). *The Analysis of Variance*. Wiley, New York.

Sellke, T. and Siegmund, D. (1983). Sequential analysis of the proportional hazards model. *Biometrika*, pages 315–326.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer-Verlag, New York.

Sheng, X. (2002). Asymptotic normality of semiparametric and nonparametric posterior distributions. *Journal of the American Statistical Association*, 97:222–235.

Skorohod, A. V. (1976). On a representation of random variables. *Theory of Probability and its Applications*, 21:628–632.

Slud, E. V. (1984). Sequential linear rank tests for two-sample censored survival data. *Annals of Statistics*, 12:551–571.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, 64:479–498.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B*, 66:187–205.

Strassen, V. (1964). An invariance principle for the law of the iterated logarithm. *Probability Theory and Related Fields*, 3:211–226.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York.

van de Geer, S. A. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press, New York.

van de Geer, S. A. (2001). Least squares estimation with complexity penalties. *Mathematical Methods of Statistics*, 10:355–374.

van der Laan, M. J. and Bryan, J. (2001). Gene expression analysis with the parametric bootstrap. *Biostatistics*, 2:445–461.

van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag, New York.

van der Vaart, A. W. (1996). Efficient maximum likelihood estimation in semiparametric mixture models. *Annals of Statistics*, 24:862–878.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, New York.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications in Statistics*. Springer-Verlag, New York.

van der Vaart, A. W. and Wellner, J. A. (2000). Preservation theorems for glivenko-cantelli and uniform glivenko-cantelli classes. In *Progress in Probability, 47*, pages 115–133. High dimensional probability, II (Seattle, WA, 1999), Birkhäuser, Boston.

Vardi, Y. (1985). Empirical distributions in selection bias models. *Annals of Statistics*, 13:178–203.

Wahba, G. (1990). *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM (Society for Industrial and Applied Mathematics), Philadelphia.

Wei, L. J. (1978). The adaptive biased coin design for sequential experiments. *Annals of Statistics*, 6:92–100.

Wellner, J. A. and Zhan, Y. H. (1996). Bootstrapping z-estimators. Technical Report 308, Department of Statistics, University of Washington.

Wellner, J. A., Zhang, Y., and Liu, H. (2002). Two semiparametric estimation methods for panel count data. Unpublished manuscript.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In Bernardo, J. M., Bayarri, M. J., Dawid, A. P., Berger, J. O., Heckerman, D., Smith, A. F. M., and West, M., editors, *Bayesian Statistics* **7**, pages 733–742. Oxford University Press, Oxford.

Wu, W. B. (2003). Empirical processes of long-memory sequences. *Bernoulli*, 9:809–831.

Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *Annals of Probability*, 22:94–116.

Zhang, D. (2001). Bayesian bootstraps for U-processes, hypothesis tests and convergence of dirichlet u-processes. *Statistica Sinica*, 11:463–478.

Zhao, L. P., Prentice, R. L., and Self, S. G. (1990). Multivariate mean parameter estimation by using a partly exponential model. *Journal of the Royal Statistical Society, Series B*, 54:805–811.

# Author Index

# List of Symbols

$\mathcal{A}$: $\sigma$-field of measurable sets, 82
$\mathbb{A}$: tangent set, 402
$A$, $A^*$: direction, orthogonal projection vectors, 403
$\mathcal{A}^*$: $\mu$-completion of $\mathcal{A}$, 83
$B^*$: adjoint of the operator $B$, 41

$B'$: transpose of $B$, 3
$B^\perp$: orthocomplement of $B$, 331
$\mathbb{B}^*$: dual of the space $\mathbb{B}$, 328
$B(\mathbb{D}, \mathbb{E})$: space of bounded linear operators between $\mathbb{D}$ and $\mathbb{E}$, 94
$BL_1(\mathbb{D})$: space of real functions on $\mathbb{D}$ with Lipschitz norm bounded by 1, 19
$\overset{\text{as}*}{\underset{M}{\rightsquigarrow}}$: conditional convergence of bootstrap outer almost surely, 20
$\overset{\text{P}}{\underset{M}{\rightsquigarrow}}$: conditional convergence of bootstrap in probability, 19
$\mathbb{B}$: Brownian bridge, 11

$\mathbb{C}$: random variable with Chernoff's distribution, 280

$C[a, b]$: space of continuous, real functions on $[a, b]$, 23
$C(T, \rho)$: space of $\rho$-continuous functions on $T$, 87
$C^c$: complement of the set $C$, 159
$\mathcal{C}^c$: class of complements of the set class $\mathcal{C}$, 159
$\mathcal{D} \times \mathcal{E}$: class of pairwise Cartesian product sets, 159
$\mathcal{C} \sqcap \mathcal{D}$: class of pairwise intersections, 159
$\mathcal{C} \sqcup \mathcal{D}$: class of pairwise unions, 159
$C_b(\mathbb{D})$: space of bounded, continuous maps $f : \mathbb{D} \mapsto \mathbb{R}$, 14
$C_M^\alpha(\mathcal{X})$: bounded real Lipschitz continuous functions on $\mathcal{X}$, 166
$\overset{\text{as}}{\rightarrow}$: convergence almost surely, 10
$\overset{\text{as}*}{\rightarrow}$: convergence outer almost surely, 14
$\overset{\text{P}}{\rightarrow}$: convergence in probability, 14
$\rightsquigarrow$: weak convergence, 11

$\operatorname{conv}\mathcal{F}$: convex hull of a class $\mathcal{F}$, 158
$\overline{\operatorname{conv}}\mathcal{F}$: closed convex hull of a class $\mathcal{F}$, 158
$\operatorname{sconv}\mathcal{F}$: symmetric convex hull of a class $\mathcal{F}$, 158
$\overline{\operatorname{sconv}}\mathcal{F}$: symmetric closed convex hull of a class $\mathcal{F}$, 158
$\operatorname{cov}[X, Y]$: covariance of $X$ and $Y$, 11

$(\mathbb{D}, d), (\mathbb{E}, e)$: metric spaces, 13
$D[a, b]$: space of cadlag functions on $[a, b]$, 22
$\mathbb{D} \times \mathbb{E}$: Cartesian product of $\mathbb{D}$ and $\mathbb{E}$, 88
$\delta B$: boundary of the set $B$, 108
$\delta_x$: point mass at $x$ or Dirac measure, 10
$\operatorname{diam} A$: diameter of the set $A$, 166

$\overline{E}$: closure of $E$, 82
$E^\circ$: interior of $E$, 82
$E(t, \beta)$: expected Cox empirical average, 56
$E_n(t, \beta)$: Cox empirical average, 5
$\mathrm{E}_*$: inner expectation, 14
$\mathrm{E}^*$: outer expectation, 14

$F$: distribution function, 9
$F$: envelope of the class $\mathcal{F}$, 18
$\mathcal{F}, \mathcal{G}$: collections of functions, 10, 19
$\mathcal{F}_\delta$, $\mathcal{F}_\infty^2$: special modifications of the class $\mathcal{F}$, 142
$\overline{\mathcal{F}}^{(P,2)}$: $L_2(P)$-closure of $\mathcal{F}$, 172
$\overline{\operatorname{sconv}}^{(P,2)}\,\mathcal{F}$: $L_2(P)$-closure of $\operatorname{sconv}\mathcal{F}$, 172
$\dot{\mathcal{F}}$: mean-zero centered $\mathcal{F}$, 171
$\mathcal{F}_1 \vee \mathcal{F}_2$: all pairwise maximums, 142
$\mathcal{F}_1 \wedge \mathcal{F}_2$: all pairwise minimums, 142
$\mathcal{F}_1 + \mathcal{F}_2$: all pairwise sums, 142
$\mathcal{F}_1 \times \mathcal{F}_2$: all pairwise products, 142
$\mathcal{F}_1 \cup \mathcal{F}_2$: union of classes $\mathcal{F}_1$ and $\mathcal{F}_2$, 142
$\{\mathcal{F} > 0\}$: class of sets $\{x : f(x) > 0\}$ over $f \in \mathcal{F}$, 160
$\mathbb{F}_n$: empirical distribution function, 9
$\mathcal{F}_\mathcal{T}$: extraction function class, 128

$\mathbb{G}$: general Brownian bridge, 11
$\mathbb{G}_n$: empirical process, 11
$\mathbb{G}_n'$, $\mathbb{G}_n''$: multiplier bootstrap empirical process, 181
$G_n$: standardized empirical distribution function, 10

$\mathbb{H}$: Hilbert space, 326

$1\{A\}$: indicator of $A$, 4
$\langle \cdot, \cdot \rangle$: inner product, 43
$I_{\theta,\eta}$: efficient information matrix, 40

$J$: Sobolev norm, 6
$J_{[]}(\delta, \mathcal{F}, \|\cdot\|)$: bracketing integral, 17
$J_{[]}^*(\delta, \mathcal{F})$, $\tilde{J}_{[]}(\delta, \mathcal{F}, \|\cdot\|)$: modified bracketing integrals, 208, 209
$J(\delta, \mathcal{F}, \|\cdot\|)$: uniform entropy integral, 18
$J^*(\delta, \mathcal{F})$: modified uniform entropy integral, 208

$\dot{\ell}_{\theta,\eta}$: score for $\theta$ when $\eta$ is fixed, 40
$\tilde{\ell}_{\theta,\eta}$: efficient score for $\theta$, 40
$\ell(t, \theta, \eta)$: likelihood for approximately least favorable submodel, 352
$\dot{\ell}(t, \theta, \eta)$: score for approximately least favorable submodel, 352
$\ell^\infty(T)$: set of all uniformly bounded real functions on $T$, 11

$L_r$: equivalence class of $r$-integrable functions, 16
$L_2^0(P)$: mean zero subspace of $L_2(P)$, 37
$\hat{\lambda}_n$: smoothing parameter, 7
$\Lambda(t)$: integrated hazard function, 5
$\operatorname{lin}\mathbb{H}$: linear span of $\mathbb{H}$, 41
$\overline{\operatorname{lin}}\,\mathbb{H}$: closed linear span of $\mathbb{H}$, 41

$m_{\theta,\eta}(X)$: objective function for semiparametric M-estimator, 401
$\mapsto$: function specifier ("maps to"), 11
$m_X(\delta)$, $m_{\mathcal{F}}(\delta)$: modulus of continuity of process $X$, class $\mathcal{F}$, 136, 172
$M$: martingale, 42
$M_n(\theta)$: M-estimating equation, 28
$T_*$: maximal measurable minorant of $T$, 14
$T^*$: minimal measurable majorant of $T$, 14
$a \vee b$: maximum of $a$ and $b$, 19
$a \wedge b$: minimum of $a$ and $b$, 6

$N(t)$: counting process, 5
$N(T)$: null space of $T$, 94
$N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$: bracketing number, 16
$N(\epsilon, \mathcal{F}, \|\cdot\|)$: covering number, 18
$\|\cdot\|_{2,1}$: special variant of $L_2$ norm, 20
$\|\cdot\|_{P,r}$: $L_r(P)$ norm, 16
$\|\cdot\|_\infty$: uniform norm, 19
$\|\cdot\|_\psi$: Orlicz $\psi$-norm, 128
$\|\cdot\|_T$: uniform norm over $T$, 87

$\mathcal{O}$: topology, 81
$\Omega, \Omega_n$: sample spaces, 14, 107
$(\Omega, \mathcal{A}, P)$: probability space, 83
$(\Omega_n, \mathcal{A}_\setminus, \mathcal{P}_\setminus)$: sequence of probability spaces, 108
$O_P(r_n)$: rate $r_n$ asymptotic boundedness in probability, 7
$o_P(r_n)$: rate $r_n$ convergence to zero in probability, 7
$\perp$: orthogonal, 326
$\emptyset$: empty set, 82
$\otimes 2$: outer product, 56

$\Pi$: projection operator, 326
$\dot{\mathcal{P}}_P$: tangent set, 37
$\mathcal{P}$: collection of probability measures, 35
$\mathbb{P}_n$: empirical measure, 10
$\mathbb{P}_n^\circ$: symmetrized, bootstrapped empirical processes, 138, 408
$\tilde{\mathbb{P}}_n$, $\hat{\mathbb{P}}_n$: weighted empirical measures, 194, 196
$P, Q$: probability measures on underlying sample space $\mathcal{X}$, 10, 18
$\mathrm{P}_*$: inner probability, 14
$\mathrm{P}^*$: outer probability, 14
$\check{\psi}_P$: influence function, 37
$\psi(P)$: model parameter, 37
$\psi_p$: the function $x \mapsto \psi_p(x) = e^{x^p} - 1$, 129
$\tilde{\psi}_P$: efficient influence function, 38
$\Psi_n$: Z-estimating equation, 24

$R(T)$: range space of $T$, 94
$\mathbb{R}$: real numbers, 3
$\bar{\mathbb{R}}$: extended reals, 14

$\sigma(f, g)$: covariance between $f(X)$ and $g(X)$, 19
$\hat{\sigma}(f, g)$: empirical estimate of $\sigma(f, g)$, 19
$\operatorname{sign}(a)$: sign of $a$, 52

$T$: index set for a stochastic process, 9
$T_n$: estimator, 36

$UC(T, \rho)$: set of uniformly continuous functions from $(T, \rho)$ to $\mathbb{R}$, 15
$U_n(t, \beta)$: Cox empirical score, 5

# Subject Index

# Springer Series in Statistics *(continued from page ii)*

*Kosorok:* Introduction to Empirical Processes and Semiparametric Inference
*Küchler/Sørensen:* Exponential Families of Stochastic Processes
*Kutoyants:* Statistical Inference for Ergodic Diffusion Processes
*Lahiri:* Resampling Methods for Dependent Data
*Lavallée:* Indirect Sampling
*Le/Zidek:* Statistical Analysis of Environmental Space-Time Processes
*Le Cam:* Asymptotic Methods in Statistical Decision Theory
*Le Cam/Yang:* Asymptotics in Statistics: Some Basic Concepts, 2nd edition
*Liese/Miescke:* Statistical Decision Theory: Estimation, Testing, Selection
*Liu:* Monte Carlo Strategies in Scientific Computing
*Manski:* Partial Identification of Probability Distributions
*Mielke/Berry:* Permutation Methods: A Distance Function Approach, 2nd edition
*Molenberghs/Verbeke:* Models for Discrete Longitudinal Data
*Mukerjee/Wu:* A Modern Theory of Factorial Designs
*Nelsen:* An Introduction to Copulas, 2nd edition
*Pan/Fang:* Growth Curve Models and Statistical Diagnostics
*Politis/Romano/Wolf:* Subsampling
*Ramsay/Silverman:* Applied Functional Data Analysis: Methods and Case Studies
*Ramsay/Silverman:* Functional Data Analysis, 2nd edition
*Reinsel:* Elements of Multivariate Time Series Analysis, 2nd edition
*Rosenbaum:* Observational Studies, 2nd edition
*Rosenblatt:* Gaussian and Non-Gaussian Linear Time Series and Random Fields
*Särndal/Swensson/Wretman:* Model Assisted Survey Sampling
*Santner/Williams/Notz:* The Design and Analysis of Computer Experiments
*Schervish:* Theory of Statistics
*Shaked/Shanthikumar:* Stochastic Orders
*Shao/Tu:* The Jackknife and Bootstrap
*Simonoff:* Smoothing Methods in Statistics
*Song:* Correlated Data Analysis: Modeling, Analytics, and Applications
*Sprott:* Statistical Inference in Science
*Stein:* Interpolation of Spatial Data: Some Theory for Kriging
*Taniguchi/Kakizawa:* Asymptotic Theory for Statistical Inference for Time Series
*Tanner:* Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, 3rd edition
*Tillé:* Sampling Algorithms
*Tsaitis:* Semiparametric Theory and Missing Data
*van der Laan/Robins:* Unified Methods for Censored Longitudinal Data and Causality
*van der Vaart/Wellner:* Weak Convergence and Empirical Processes: With Applications to Statistics
*Verbeke/Molenberghs:* Linear Mixed Models for Longitudinal Data
*Weerahandi:* Exact Statistical Methods for Data Analysis