WILEY SERIES IN PROBABILITY AND MATHEMATICAL STATISTICS

ESTABLISHED BY WALTER A. SHEWHART AND SAMUEL S. WILKS

Editors

Vic Barnett, Ralph A. Bradley, J. Stuart Hunter, David G. Kendall, Rupert G. Miller, Jr., Stephen M. Stigler, Geoffrey S. Watson

Probability and Mathematical Statistics

ADLER • The Geometry of Random Fields
ANDERSON • The Statistical Analysis of Time Series
ANDERSON • An Introduction to Multivariate Statistical Analysis, *Second Edition*
ARAUJO and GINE • The Central Limit Theorem for Real and Banach Valued Random Variables
ARNOLD • The Theory of Linear Models and Multivariate Analysis
BARLOW, BARTHOLOMEW, BREMNER, and BRUNK • Statistical Inference Under Order Restrictions
BARNETT • Comparative Statistical Inference, *Second Edition*
BHATTACHARYYA and JOHNSON • Statistical Concepts and Methods
BILLINGSLEY • Probability and Measure
BOROVKOV • Asymptotic Methods in Queuing Theory
BOSE and MANVEL • Introduction to Combinatorial Theory
CASSEL, SARNDAL, and WRETMAN • Foundations of Inference in Survey Sampling
CHEN • Recursive Estimation and Control for Stochastic Systems
COCHRAN • Contributions to Statistics
COCHRAN • Planning and Analysis of Observational Studies
DE FINETTI • Theory of Probability, Volume II
DOOB • Stochastic Processes
EATON • Multivariate Statistics: A Vector Space Approach
FABIAN • Introduction to Probability and Mathematical Statistics
FELLER • An Introduction to Probability Theory and Its Applications, Volume I, *Third Edition*, Revised; Volume II, *Second Edition*
FULLER • Introduction to Statistical Time Series
GRENANDER • Abstract Inference
GUTTMAN • Linear Models: An Introduction
HANNAN • Multiple Time Series
HANSEN, HURWITZ, and MADOW • Sample Survey Methods and Theory, Volumes I and II
HETTMANSPERGER • Statistical Inference Based on Ranks
HOEL • Introduction to Mathematical Statistics, *Fifth Edition*
HUBER • Robust Statistics
IMAN and CONOVER • A Modern Approach to Statistics
IOSIFESCU • Finite Markov Processes and Applications
ISAACSON and MADSEN • Markov Chains
JOHNSON and BHATTACHARYYA • Statistics: Principles and Methods
LAHA and ROHATGI • Probability Theory
LARSON • Introduction to Probability Theory and Statistical Inference, *Third Edition*
LEHMANN • Testing Statistical Hypotheses
LEHMANN • Theory of Point Estimation
MATTHES, KERSTAN, and MECKE • Infinitely Divisible Point Processes
MUIRHEAD • Aspects of Multivariate Statistical Theory
PARZEN • Modern Probability Theory and Its Applications
PURI and SEN • Nonparametric Methods in Multivariate Analysis
RANDLES and WOLFE • Introduction to the Theory of Nonparametric Statistics
RAO • Linear Statistical Inference and Its Applications, *Second Edition*
RAO and SEDRANSK • W.G. Cochran's Impact on Statistics
ROHATGI • An Introduction to Probability Theory and Mathematical Statistics

Probability and Mathematical Statistics (Continued)

ROHATGI • Statistical Inference
ROSS • Stochastic Processes
RUBINSTEIN • Simulation and The Monte Carlo Method
SCHEFFE • The Analysis of Variance
SEBER • Linear Regression Analysis
SEBER • Multivariate Observations
SEN • Sequential Nonparametrics: Invariance Principles and Statistical Inference
SERFLING • Approximation Theorems of Mathematical Statistics
TJUR • Probability Based on Radon Measures
WILLIAMS • Diffusions, Markov Processes, and Martingales, Volume I: Foundations
ZACKS • Theory of Statistical Inference

Applied Probability and Statistics

ABRAHAM and LEDOLTER • Statistical Methods for Forecasting
AGRESTI • Analysis of Ordinal Categorical Data
AICKIN • Linear Statistical Analysis of Discrete Data
ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE, and WEISBERG • Statistical Methods for Comparative Studies
ARTHANARI and DODGE • Mathematical Programming in Statistics
BAILEY • The Elements of Stochastic Processes with Applications to the Natural Sciences
BAILEY • Mathematics, Statistics and Systems for Health
BARNETT • Interpreting Multivariate Data
BARNETT and LEWIS • Outliers in Statistical Data, *Second Edition*
BARTHOLOMEW • Stochastic Models for Social Processes, *Third Edition*
BARTHOLOMEW and FORBES • Statistical Techniques for Manpower Planning
BECK and ARNOLD • Parameter Estimation in Engineering and Science
BELSLEY, KUH, and WELSCH • Regression Diagnostics: Identifying Influential Data and Sources of Collinearity
BHAT • Elements of Applied Stochastic Processes, *Second Edition*
BLOOMFIELD • Fourier Analysis of Time Series: An Introduction
BOX • R. A. Fisher, The Life of a Scientist
BOX and DRAPER • Evolutionary Operation: A Statistical Method for Process Improvement
BOX, HUNTER, and HUNTER • Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building
BROWN and HOLLANDER • Statistics: A Biomedical Introduction
BROWNLEE • Statistical Theory and Methodology in Science and Engineering, *Second Edition*
CHAMBERS • Computational Methods for Data Analysis
CHATTERJEE and PRICE • Regression Analysis by Example
CHOW • Analysis and Control of Dynamic Economic Systems
CHOW • Econometric Analysis by Control Methods
COCHRAN • Sampling Techniques, *Third Edition*
COCHRAN and COX • Experimental Designs, *Second Edition*
CONOVER • Practical Nonparametric Statistics, *Second Edition*
CONOVER and IMAN • Introduction to Modern Business Statistics
CORNELL • Experiments with Mixtures: Designs, Models and The Analysis of Mixture Data
COX • Planning of Experiments
DANIEL • Biostatistics: A Foundation for Analysis in the Health Sciences, *Third Edition*
DANIEL • Applications of Statistics to Industrial Experimentation
DANIEL and WOOD • Fitting Equations to Data: Computer Analysis of Multifactor Data, *Second Edition*
DAVID • Order Statistics, *Second Edition*
DAVISON • Multidimensional Scaling
DEMING • Sample Design in Business Research
DILLON and GOLDSTEIN • Multivariate Analysis: Methods and Applications

continued on back

Nonparametric Density Estimation

THE L_1 VIEW

Luc Devroye

McGill University
Montreal, Canada

László Györfi

Hungarian Academy of Sciences
Budapest, Hungary

John Wiley & Sons

New York • Chichester • Brisbane • Toronto • Singapore

Library of Congress Cataloging in Publication Data:

Devroye, Luc.
Nonparametric density estimation.

(Wiley series in probability and mathematical statistics. Probability and mathematical statistics)
Includes index.
1. Distribution (Probability theory) 2. Estimation theory. 3. Nonparametric satistics. I. Györfi, László. II. Title. III. Series.

QA273.6.D48 1984 519.2 84-15198
ISBN 0-471-81646-9

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

TO BEA AND KATI

Preface

This is a book about the L_1 convergence of density estimates that are based on a sample of R^d-valued independent identically distributed random vectors. In it, we try to develop a smooth L_1 theory because the better studied L_2 theory has led to various anomalies and misconceptions. The book is not an exhaustive description of all known L_1 results, but rather a collection of observations with emphasis on those results that lead to a better understanding of density estimation. Of course, by intentionally limiting ourselves to the L_1 theory, we are omitting some interesting and often profound work on nonparametric density estimation.

Although we hope that this book is entertaining in places, most of it, in fact, is rather dull except perhaps to the odd technical fanatic. Thus, we do not recommend it for class notes or for reading during TV commercials. We had to make a few sacrifices for the goals that we set ourselves—conciseness, generality, and optimality. For example, shallow results padded with unnecessary conditions, as a rule, have simple and short proofs. To generalize the results and get rid of the convenient conditions, sometimes long devious paths must be followed. In doing so, one often stumbles on nice tangential results worth reporting, and before one realizes it, the whole enterprise becomes a nearly impenetrable technical jungle. The book grew out of excitement and enthusiasm: excitement every time one of us closed a gap or crossed a bridge, and enthusiasm about simple things such as beautiful inequalities. Our excitement and haste are thus to blame for any errors that the reader may discover.

In our choice of topics and selection of mathematical tools, we were influenced by the original papers on nonparametric density estimation (Parzen, Rosenblatt), by the modern French school (Geffroy, Bosq, Deheuvels, Abou-Jaoude, Bretagnolle, Huber, Birgé), and by some scattered relatively recent work on related topics (Geman, Steele, Stone). We would like to thank the people who have directly helped us through discussions and lectures: Terry Wagner, Sandor Csibi, Clark Penrod, Paul Deheuvels,

Adam Krzyzak, Peter Hall, and Godfried Toussaint. We would also like to thank NSERC Canada for its generous grant support and McGill University for not taking any overhead charges from this grant. Finally, we would like to thank a number of colleagues and friends who, often without realizing it, have contributed to our understanding of density estimation via personal discussions: Alain Berlinet, Lucien Birgé, Denis Bosq, Jean Bretagnolle, Pat Brockett, Gérard Collomb, Tom Cover, Ben Fox, Stuart Geman, Piet Groeneboom, Wilfrid Grossmann, Antonio Gualtierotti, Catherine Huber, Jean-Pierre Lecoutre, Fred Machell, Manny Parzen, Georg Pflug, Pal Révész, David Scott, Mike Steele, Jim Thompson, and Wolfgang Wertz.

LUC DEVROYE
LÁSZLÓ GYÖRFI

Montreal, Canada
Budapest, Hungary
October 1984

Contents

CHAPTER 1

Introduction

There is a vast literature on nonparametric density estimation, and any book on this topic is necessarily of limited scope. This book is no exception. Our selection, we were guided by general principles: for example, we stubbornly treat all densities as members of L_1 and not of L_2 or L_∞ as is done elsewhere. We also do not cover estimates that are not densities since we believe that a density should be estimated by a density. Because L_1 is the natural space for densities, an in-depth treatment of its properties leads to a very smooth theory, uncluttered by unnecessary conditions. We will try to state all our theorems in their most general (simplest) form.

This book deals with the following problem: we are given *data* $X_1, \ldots, X_n$, independent identically distributed random vectors taking values in R^d and having a common density f. A *density estimate* is a sequence $f_1, f_2, \ldots$, where for each n, $f_n(x) = f_n(x; X_1, \ldots, X_n)$ is a real-valued Borel measurable function of its arguments, and for fixed n, $X_1, \ldots, X_n$, f_n is a density on R^d.

Our choice of the L_1 distance $J_n = \int |f_n - f|$ is motivated by its invariance under monotone transformations of the coordinate axes and the fact that it is always well-defined. Consider, for example, two random vectors X and Y with densities f and g, respectively. Now apply the transformation T: $R^d \to R^d$ to X and Y, where T is sufficiently rich, that is, if $\mathscr{B}$ is the class of all Borel sets of R^d, then $\{T^{-1}B | B \in \mathscr{B}\} = \mathscr{B}$ (this implies that the transformation is one-to-one). The densities of $T(X)$ and $T(Y)$ are f^* and g^*, but regardless of T we have

$$\int |f - g| = \int |f^* - g^*|. \tag{1}$$

In particular, for $d = 1$, J_n is invariant under continuous strictly monotone transformations. Property (1) is a corollary of the following theorem:

THEOREM 1 (Scheffé, 1947). *For all densities f and g on R^d,*

$$\int |f - g| = 2 \sup_{B \in \mathscr{B}} \left| \int_B f - \int_B g \right|. \tag{2}$$

Proof. Let $B = \{f > g\}$, and let $A \in \mathscr{B}$. Because $\int (f - g) = 0$, $\int |f - g| = 2 \int_B (f - g)$, and, thus, (2) follows with " $\leq$ " instead of " $=$ ". Also,

$$\begin{aligned} \left| \int_A f - \int_A g \right| &= \left| \int_{A \cap B} (f - g) + \int_{A \cap B^c} (f - g) \right| \\ &\leq \max\left(\int_{A \cap B} (f - g), \int_{A \cap B^c} (g - f) \right) \\ &\leq \max\left(\int_B (f - g), \int_{B^c} (g - f) \right) = \frac{1}{2} \int |f - g|, \quad \text{all } A \in \mathscr{B}, \end{aligned}$$

where B^c denotes the complement of B. This completes the proof of (2).

The invariance property (1) follows easily:

$$\begin{aligned} \int |f^* - g^*| &= 2 \sup_B \left| \int_B f^* - \int_B g^* \right| = 2 \sup_B \left| \int_{T^{-1}B} f - \int_{T^{-1}B} g \right| \\ &= 2 \sup_B \left| \int_B f - \int_B g \right| = \int |f - g|. \end{aligned}$$

In other words, when $d = 1$, we can draw the graphs of f and g on any linear or nonlinear scale of our choice, or even consider transforms T: $R \to [0, 1]$ and draw the transformed densities conveniently on $[0, 1]$, and get a visual idea of the size of J_n by taking a quick look at the size of the area lying between the graphs of the densities. Also, Theorem 1 relates the L_1 distance between f and g to the maximal error committed if we were to estimate the probabilities of all the Borel sets using f and g, respectively.

Consider now the L_p distance $(\int |f - g|^p)^{1/p}$, and replace X and Y by aX and aY where $a \neq 0$ is a scale factor, and our dimension is 1. Thus, the density of aX is $f^*(x) = (1/a) f(x/a)$. Therefore,

$$\left(\int |f^* - g^*|^p \right)^{1/p} = a^{(1-p)/p} \left(\int |f - g|^p \right)^{1/p}. \tag{3}$$

Except for $p = 1$, all L_p distances depend upon the scale that is used. They cannot be compared with each other on a universal scale, such as the one provided by (2): for example, when a density estimate is used to estimate f,

and is then used to estimate g, the comparison between $\int (f_n - f)^2$ and $\int (f_n - g)^2$ is meaningless, because of (3). Yet, by Scheffé's Theorem, $\int |f_n - f|$ and $\int |f_n - g|$ can be compared in an absolute manner. It is thus conceivable to declare that a given density f is easier to estimate with a given f_n than g.

The reader will have no difficulty with the proof of the following statement: *for any f and any $p > 1$, there exist sequences of densities f_n* and g_n such that $\int |f_n - f| \downarrow 0$, $\int |f_n - f|^p \uparrow \infty$, $\int |g_n - f| = c > 0$, and $\int |g_n - f|^p \downarrow 0$. Thus, simple relations or inequalities between the L_1 distance and the L_p distance do not exist.

Density estimates are all based upon the Lebesgue density theorem: when S_{xh} is the closed sphere of radius h centered at x, and λ is Lebesgue measure, then

$$\lim_{h \downarrow 0} \int_{S_{xh}} \frac{f(y)\,dy}{\lambda(S_{xh})} = f(x), \quad \text{almost all } x. \tag{4}$$

The quantity on the left-hand side of (4) is $P(X_1 \in S_{xh})/\lambda(S_{xh})$ and can thus be approximated by

$$f_n(x) = \sum_{i=1}^{n} \frac{I_{[X_i \in S_{xh}]}}{n\lambda(S_{xh})} \tag{5}$$

where I is the indicator function. Estimate (5) was suggested by Rosenblatt in 1956. For a good approximation in (4) it is necessary that h be small. Yet, when h is small, the variance of (5) increases because fewer points are expected to fall in S_{xh}. In the choice of h, one must balance both effects, and this creates interesting theoretical problems.

In Chapter 2, general approximation theorems of the type (4) are presented. In Chapters 3, 5, and 6, two estimates are considered in parallel, the *kernel estimate* and the *histogram estimate*. In particular, we give necessary and sufficient conditions on h for all types of convergence of J_n (Chapter 3), rate of convergence results for $E(J_n)$ featuring a universal lower bound for $\liminf_{n \to \infty} n^{2/5} E(J_n)$ for all kernel estimates and all densities f (Chapter 5), and convergence theorems for kernel estimates in which h is chosen as a function of the data (Chapter 6).

In Chapter 4, we show that for all density estimates, $E(J_n)$ can be forced to tend to 0 at any prespecified rate merely by choosing f in an appropriate class of densities such as the class of all infinitely many-times differentiable densities, or the class of all densities with support in $[0,1]^d$ bounded by 2. Thus, there does not exist an estimate, however sophisticated, for which $E(J_n)$ decreases at some given rate for all f. For the study of rates of

convergence of $E(J_n)$ we have to impose conditions on f, and by the results of Chapter 4, it is clear that tail conditions alone, or smoothness conditions alone, are not sufficient. The remaining chapters illustrate the basic theory.

Chapters 7 through 12 can be read in any order and have varying levels of sophistication depending upon the intended readership. In Chapter 9, we discuss the transformed kernel estimate. In Chapter 12, several estimates related to orthogonal series expansions are given. Other estimates are described in Chapter 7. Chapters 8, 10, and 11 are more application-oriented. In Chapter 8, for example, we tackle the problem of the use of f_n in simulations.

In Chapter 10, we show that every density estimate has its analogue in discrimination, and that there is a close connection between the probability of error in discrimination and J_n. Finally, in Chapter 11, we consider among other things some applications in detection theory.

Many topics are not covered, and many questions are left unanswered. The most important omissions include an asymptotic distribution theory for J_n, a law of the iterated logarithm for J_n, results about the rate of convergence of $E(J_n)$ in higher dimensions, methods for estimating J_n, and confidence intervals for J_n.

Each chapter has its own list of references. Additional references about other properties of the estimates treated here (such as their behavior when $X_1, \ldots, X_n$ are not independent; or L_p properties for $1 < p \leq \infty$; or laws of the iterated logarithm) or about other estimates can be found in the monographs of Wertz (1978), Tapia and Thompson (1978), Nadaraya (1983), Prakasa Rao (1983), and in the survey papers and bibliographies of Cover (1972), Fryer (1977), Földes and Révész (1974), Leonard (1978), Révész (1972), Tarter and Kronmal (1976), Wegman (1972a, 1972b), Wertz and Schneider (1979) and Bean and Tsokos (1980).

Within each chapter, the formulas are numbered (1), (2), (3), and so on, and the theorems are numbered 1, 2, 3, and so on. When we refer to Theorem 3 within a chapter, we mean Theorem 3 of the same chapter. Otherwise, we will add the chapter's number as in Theorem 2.3. The chapters have the following dependence structure: 2 is necessary for 3, 4, and 6; 3 is a prerequisite for 7 and 10; 4 is needed for 5; and 5 in turn is a prerequisite for 8 and 9.

REFERENCES

S. J. Bean and C. P. Tsokos (1980). Developments in nonparametric density estimation, *International Statistical Review* **48**, pp. 267–287.

T. M. Cover (1972). A hierarchy of probability density function estimates, in *Frontiers of Pattern Recognition*, Academic Press, New York, pp. 83–98.

A. Földes and P. Révész (1974). A general method for density estimation, *Studia Scientiarium Mathematicarum Hungarica* **9**, pp. 81–92.

M. J. Fryer (1977). A review of some nonparametric methods of density estimation, *Journal of the Institute of Mathematics and Applications*, **20**, pp. 335–354.

T. Leonard (1978). Density estimation, stochastic processes and prior information, *Journal of the Royal Statistical Society B* **40**, pp. 113–146.

E. A. Nadaraya (1983). *Nonparametric Estimation of Probability Density and Regression Curve*, The Publishing Office of Tbilisi University, Tbilisi, USSR (in Russian).

B. L. S. Prakasa Rao (1983). *Nonparametric Functional Estimation*, Academic Press, New York.

P. Révész (1972). On empirical density function, *Periodica Mathematica Hungarica* **2**, pp. 85–110.

M. Rosenblatt (1956). Remarks on some nonparametric estimates of a density function, *Annals of Mathematical Statistics* **27**, pp. 832–835.

H. Scheffé (1947). A useful convergence theorem for probability distributions, *Annals of Mathematical Statistics* **18**, pp. 434–458.

R. A. Tapia and J. R. Thompson (1978). *Nonparametric Probability Density Estimation*, The Johns Hopkins University Press, Baltimore.

M. E. Tarter and R. A. Kronmal (1976). An introduction to the implementation and theory of nonparametric density estimation, *The American Statistician* **30**, pp. 105–112.

E. J. Wegman (1972a). Nonparametric probability density estimation. I. A summary of available methods, *Technometrics* **14**, pp. 533–546.

E. J. Wegman (1972b). Nonparametric probability density estimation. II. A comparison of density estimation methods, *Journal of Statistical Computation and Simulation* **1**, pp. 225–245.

W. Wertz (1978). *Statistical Density Estimation. A Survey*, Vandenhoeck and Ruprecht, Gottingen, Applied Statistics and Econometrics Series 13.

W. Wertz and B. Schneider (1979). Statistical density estimation: A bibliography, *International Statistical Review* **47**, pp. 155–175.

CHAPTER 2

Differentiation of Integrals

The most important tool in our analysis is Lebesgue's density theorem (1.4). All the results of a similar type are collected in this chapter. For proofs and illuminating discussions, we refer the reader to Chapters 7 and 9 of Wheeden and Zygmund (1977) and Chapters 1–3 of de Guzman (1975). See also Shapiro (1969), Stein (1970), Hayes and Pauc (1970) and de Guzman (1981). Throughout this section, λ is Lebesgue measure on R^d, K is a Borel measurable function on R^d, f is a density on R^d, $h > 0$ is a positive number, $K_h(x) = (1/h^d)K(x/h)$, and " $*$ " is the convolution operator, for example, when $K \in L_1(\lambda)$,

$$f * K(x) = \int f(y)K(x-y)\,dy = \int K(y)f(x-y)\,dy.$$

THEOREM 1. *For all functions* $f, g \in L_1(\lambda)$, $\int |f * g| \le \int |f| \int |g|$ (*Young's inequality*). *For all* $f, K \in L_1(\lambda)$ *with* $\int K = 1$, *we have*

$$\lim_{h \downarrow 0} \int |f * K_h - f| = 0.$$

Proof. The first inequality follows by a change in the order of integration (which is justified for nonnegative integrands):

$$\int \left| \int f(y)g(x-y)\,dy \right| dx \le \int \int |f(y)||g(x-y)|\,dy\,dx$$

$$= \int |f(y)| \int |g(x-y)|\,dx\,dy = \int |g| \int |f|.$$

We first prove the second statement of Theorem 1 for continuous f vanishing outside a compact set. Let $\omega(t)$ be the modulus of continuity of f, $\omega(t) = \sup_{\|x-y\| \le t} |f(x) - f(y)|$, and let m be a large number. Split K into $K' + K''$ where $K' = KI_{[\|x\| \le M]}$, $K'' = KI_{[\|x\| > M]}$. Clearly,

$$\int |f * K_h - f| \le \int \left| f * K_h' - f\left(\int K_h'\right)\right| + \int |f * K_h''| + \int f \int |K_h''|. \quad (1)$$

The last two terms of (1) do not exceed $2\int |K_h''| = 2\int |K''|$, which can be made as small as desired by choice of M. The first term on the right-hand side of (1) is $o(1)$ because it equals, for some large compact set A,

$$\begin{aligned}\int_A \left| f * K_h' - f\left(\int K_h'\right)\right| &\le \int_A \int |f(x-y) - f(x)||K_h'(y)|\,dy\,dx \\ &\le \omega(Mh) \int_A \int |K_h'(y)|\,dy\,dx \\ &\le \omega(Mh)\lambda(A)\int |K| = o(1).\end{aligned}$$

For all f, and all continuous g with compact support, we have

$$\begin{aligned}\int |f * K_h - f| &\le \int |f - g| * |K_h| + \int |f - g| + \int |g * K_h - g| \\ &\le \left(\int |K| + 1\right)\int |f - g| + o(1),\end{aligned}$$

and this can be made as small as desired by choice of g.

THEOREM 2 (Lebesgue Density Theorem). *Let $\mathscr{B}$ be a class of Borel sets of R^d having the following property*:

$$\sup_{B \in \mathscr{B}} \frac{\lambda(\textit{smallest cube centered at } 0 \textit{ containing } B)}{\lambda(B)} < \infty.$$

Then, for any sequence of sets B_k from $\mathscr{B}$ with $\lambda(B_k) \to 0$, we have

$$\lim_{k \to \infty} \int_{x + B_k} \frac{|f(y) - f(x)|\,dy}{\lambda(B_k)} = 0, \quad \textit{almost all } x. \quad (2)$$

Thus also,

$$\lim_{k\to\infty}\int_{x+B_k}\frac{f(y)\,dy}{\lambda(B_k)}=f(x),\quad \textit{almost all } x. \tag{3}$$

The points for which (2) *and* (3) *are valid are called* Lebesgue points *for* f. *The set of Lebesgue points depends only upon* f.

For the proof of Theorem 2, see Wheeden and Zygmund (1977, pp. 108–109). It is noteworthy that for $\mathscr{B}$ we can take all spheres centered at the origin (in which case we obtain the classical version of the Lebesgue density theorem), or all sets of the form aA where $a > 0$ and A is a fixed compact set of R^d, but that we cannot take all rectangles containing the origin.

THEOREM 3. *Let* $K \in L_1(\lambda)$ *with* $\int K = 1$. *Assume that* K *has an integrable radial majorant* $\psi \in L_1(\lambda)(\psi(x) = \sup_{\|y\|\geq\|x\|}|K(y)|)$. *Then*

$$f * K_h \to f \quad \textit{as } h \downarrow 0 \quad \textit{for almost all } x.$$

Theorem 3 is due to Stein (1970, pp. 62–63). It is, for example, sufficient that K is bounded, in $L_1(\lambda)$, $\int K = 1$ and $K(x) \leq a/\|x\|^{d+\varepsilon}$ for some $\varepsilon > 0$, $a > 0$. This is the version found in Wheeden and Zygmund (1977, pp. 152–153). Of course, for bounded K with compact support, Theorem 3 is a simple corollary of Theorem 2.

Theorem 4 is the converse of Theorem 1.

THEOREM 4. *Let* K *be a density on* R^d. *Then* $\int|f * K_h - f| > 0$ *for all* $h > 0$, *and when* $h = h_n$ *is a sequence of numbers,* $\lim_{n\to\infty}\int|f * K_h - f| = 0$ *implies* $h \to 0$.

Proof. Let ϕ and ψ be the characteristic functions of f and K, respectively. Thus, $f * K_h$ has characteristic function $\psi(ht)\phi(t)$, $t \in R^d$. Clearly, $\int|f * K_h - f| = 0$ implies that $f = f * K_h$ for almost all x, and thus that $\phi(t) = \phi(t)\psi(ht)$ for all $t \in R^d$. For $\phi(t) \neq 0$, that is, at least in a neighborhood of the origin, $\psi(ht) = 1$. But since $h \neq 0$, this implies that ψ cannot be the characteristic function of a density on R^d, and we have a contradiction. This proves the first part of Theorem 4.

To prove the second statement of Theorem 4, we assume first that $\lim h = \infty$. By Fatou's lemma, $\int|f * K_h - f| \to 0$ implies $\liminf|f * K_h - f| = 0$, for almost all x. But since $f * K_h \to 0$ for almost all x, we must have $f = 0$ for almost all x, and this is impossible. Assume next that $\lim h = c \in (0, \infty)$. Clearly, $\int|f * K_h - f| \geq \int|f * K_c - f| - \int|f * K_c - f * K_h|$. By the

first part of the theorem, it suffices to show that $\int |f * K_c - f * K_h| \to 0$ to reach a contradiction, thereby concluding the proof of Theorem 4. Let K' be a continuous function with compact support. By Theorem 1, and the Lebesgue dominated convergence theorem,

$$\begin{aligned}\int |f * K_c - f * K_h| &\le \int |K_c - K_h| \\ &\le \int |K_c - K_c'| + \int |K_c' - K_h'| + \int |K_h' - K_h| \\ &= 2\int |K - K'| + \int |K_c' - K_h'| \\ &= 2\int |K - K'| + o(1). \end{aligned} \tag{4}$$

The last expression in the chain of inequalities (4) can be made small by choosing K' close enough to K in $L_1(\lambda)$.

To study the histogram estimates, we need some martingale convergence theorems. Consider a sequence of partitions $\mathscr{P}_n = \{A_{nj},\ j = 1, 2, \dots\}$, $n \ge 1$ with $\lambda(A_{nj}) \in (0, \infty)$ for all n, j, and all A_{nj} are Borel sets of R^d. The sequence is said to be *nested* when $\mathscr{P}_{n+1}$ is a refinement of $\mathscr{P}_n$ for all n. It is called *cubic* if there exist positive constants $a_1, \dots, a_d$ and a sequence of positive numbers $h = h_n$ such that each A_{nj} is of the form $\prod_{i=1}^d [a_i k_i h, a_i(k_i + 1)h)$ where $k_1, \dots, k_d$ are integers. In what follows, we let $\mathscr{B}_n = \sigma(\mathscr{P}_n)$ be the σ-algebra generated by $\mathscr{P}_n$, $\mathscr{B}_n' = \sigma(\cup_{m \ge n} \mathscr{P}_m)$, and $\mathscr{B}$ = class of all Borel sets of R^d. Throughout, we assume that

$$\mathscr{B} = \bigcap_{n=1}^{\infty} \mathscr{B}_n'. \tag{5}$$

Condition (5) states that the sequence of partitions must be rich enough.

Consider the function

$$g_n(x) = \int_{A_{nj}} \frac{f}{\lambda(A_{nj})}, \qquad x \in A_{nj}. \tag{6}$$

For sequences of partitions satisfying (5), Abou-Jaoude (1976) has proved the following strong analogue of Theorems 1 and 4.

THEOREM 5 (Abou-Jaoude, 1976).

$$\int |g_n - f| \to 0 \quad \textit{for all densities } f$$

if and only if for all $A \in \mathcal{B}$ *with* $0 < \lambda(A) < \infty$ *and for all* $\varepsilon > 0$ *there exists an* n_0 *such that for all* $n \geq n_0$ *we can find an* A_n *in* $\mathcal{B}_n$ *with* $\lambda(A \Delta A_n) < \varepsilon$ (*here* Δ *denotes the symmetric difference*).

Theorem 5 is proved in Abou-Jaoude (1976, pp. 216–219).

THEOREM 6. *For a cubic sequence of partitions,* (5) *is satisfied and* $\int |g_n - f| \to 0$ *for all densities* f *if and only if* $\lim h = 0$.

Proof. First, it is clear that $\lim h = 0$ is necessary and sufficient for (5). For example, the sufficiency follows from the observation that $\cap_{n=1}^{\infty} \mathcal{B}_n'$ contains all sets of the form $\prod_{i=1}^{d} (-\infty, x_i)$ for all $x = (x_1, \ldots, x_d) \in R^d$, and that these sets generate the Borel σ-algebra.

Thus, we will just check the condition of Theorem 5. Because λ is a regular measure on R^d (i.e., all Borel sets are decreasing limits of open sets), we should only consider bounded open sets O. But every set O is a countable union of rectangles. Thus, for every $\varepsilon > 0$ there exists a finite collection of rectangles $R_1, \ldots, R_N$ such that $\lambda(O - \cup_{i=1}^{N} R_i) < \varepsilon$. Thus, it suffices to establish the condition of Theorem 5 for a finite number of rectangles, and, in fact, for one rectangle. But for one rectangle, the condition is trivially satisfied.

We should mention here that for cubic sequences of partitions, nested or not, $g_n \to f$ for almost all x, by a trivial application of Theorem 2.

We have seen that pointwise convergence theorems usually require more conditions than integral convergence theorems, for example, compare Theorem 3 with Theorem 1. This is because pointwise convergence is a concept that is strictly stronger than L_1 convergence:

THEOREM 7 (Scheffé, 1947). *Let* f_n *be a sequence of densities on* R^d *tending almost everywhere to a density* f. *Then* $\int |f_n - f| \to 0$.

Proof. By Theorem 1.1, $\int |f_n - f| = 2 \int_{f \geq f_n} (f - f_n) \to 0$, where we used the Lebesgue dominated convergence theorem.

THEOREM 8 (Glick, 1974). *Let* f_n *be a density estimate on* R^d, *and let* f *be a density on* R^d. *If* $f_n \to f$ *in probability as* $n \to \infty$, *for almost all* x, *then* $\int |f_n - f| \to 0$ *in probability* (*and thus in the mean*) *as* $n \to \infty$. *If* $f_n \to f$ *almost surely as* $n \to \infty$, *for almost all* x, *then* $\int |f_n - f| \to 0$ *almost surely as* $n \to \infty$.

Proof. We will write $(\ )_+$ for the positive part of a function. By assumption, $(f - f_n)_+ \to 0$ in probability for almost all x. Since $(f - f_n)_+ \le f$, we thus have $E((f - f_n)_+) \to 0$ for almost all x, by the Lebesgue dominated convergence theorem. But by another application of the Lebesgue dominated convergence theorem, $E(\int |f_n - f|) = E(2\int (f - f_n)_+) = 2\int E((f - f_n)_+) \to 0$.

For the second part of the theorem, we let $(\Omega, \mathcal{F}, P)$ be the probability space of $X_1, X_2, \ldots$, with probability element ω. By Fubini's theorem,

$$P(\omega\colon f_n(x) \nrightarrow f(x)) = 0 \quad \text{for almost all } x(\lambda)$$

if and only if

$$\{(\omega, x)\colon f_n(x) \nrightarrow f(x)\} \quad \text{has } P \times \lambda \text{ measure } 0$$

if and only if

$$\lambda(x\colon f_n(x) \nrightarrow f(x)) = 0 \quad \text{for almost all } \omega(P).$$

Let Ω' be the last set of ω's. By the Lebesgue dominated convergence theorem, $\int |f_n - f| \to 0$ for all $\omega \in \Omega'$. The theorem now follows since $P(\Omega') = 1$.

REFERENCES

S. Abou-Jaoude (1976). Conditions nécessaires et suffisantes de convergence L_1 en probabilité de l'histogramme pour une densité, *Annales de l'Institut Henri Poincaré* **12**, pp. 213–231.

M. de Guzman (1975). *Differentiation of Integrals in R^n*, Lecture Notes in Mathematics # 481, Springer-Verlag, Berlin.

M. de Guzman (1981). *Real Variable Methods in Fourier Analysis*, North-Holland, Amsterdam.

L. Devroye (1983). The equivalence of weak, strong and complete convergence in L_1 for kernel density estimates, *Annals of Statistics* **11**, pp. 896–904.

N. Glick (1974). Consistency conditions for probability estimators and integrals of density estimators, *Utilitas Mathematica* **6**, pp. 61–74.

C. A. Hayes and C. Y. Pauc (1970). *Derivation and Martingales*, Springer-Verlag, New York.

J. Neveu (1975). *Discrete-Parameter Martingales*, North Holland, Amsterdam.

H. Scheffé (1947). A useful convergence theorem for probability distributions, *Annals of Mathematical Statistics* **18**, pp. 434–458.

H. S. Shapiro (1969). *Smoothing and Approximation of Functions*, Van Nostrand Reinhold, New York.

E. M. Stein (1970). *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, New Jersey.

R. L. Wheeden and A. Zygmund (1977). *Measure and Integral*, Marcel Dekker, New York.

CHAPTER 3

Consistency

1. KERNEL ESTIMATE

The *kernel estimate* (Parzen, 1962; Rosenblatt, 1956; Cacoullos, 1966) is defined by

$$f_n(x) = (nh^d)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where $h = h_n$ is a sequence of positive numbers, and K is a Borel measurable function (kernel) satisfying $K \geq 0$, $\int K = 1$. The main result of this section is that for the kernel estimate all types of convergence to 0 for J_n are equivalent. Theorem 1 given below states that either $J_n \to 0$ completely for all f, or J_n does not converge to 0 in probability for a single f. There is no intermediate situation. A weak analogue of Theorem 1 for histogram estimates is given in Section 3. Theorem 1 was first published in Devroye (1983), but some key ideas go back to Abou-Jaoude (1977).

THEOREM 1. *Let K be a nonnegative Borel measurable function on R^d with $\int K = 1$. Then the following statements are equivalent:*

(i) $J_n \to 0$ *in probability as* $n \to \infty$, *some* f.

(ii) $J_n \to 0$ *in probability as* $n \to \infty$, *all* f.

(iii) $J_n \to 0$ *almost surely as* $n \to \infty$, *all* f.

(iv) $J_n \to 0$ *exponentially as* $n \to \infty$ (*i.e., for all* $\varepsilon > 0$, *there exist* $r, n_0 > 0$ *such that* $P(J_n \geq \varepsilon) \leq e^{-rn}, n \geq n_0$), *all* f.

(v) $\lim_{n \to \infty} h = 0$, $\lim_{n \to \infty} nh^d = \infty$.

In (iv), r can be chosen independently of f. Also, (v) implies (iv) when K is merely absolutely integrable and $\int K = 1$.

REMARK 1. We will show that (v) implies that $P(J_n \geq \varepsilon) \leq e^{-rn\varepsilon^2}$ for all $\varepsilon \in (0,1)$ and all $n \geq n_0$, where n_0 depends upon f and ε. For fixed f,

there exist functions $h_0(\varepsilon)$ and $c_0(\varepsilon)$ such that for

$$\left(\frac{c_0(\varepsilon)}{n}\right)^{1/d} < h < h_0(\varepsilon),$$

this exponential bound is valid. Thus, for a given ε, h may remain constant and the exponential inequality remains valid nevertheless.

3 PROOF OF THEOREM 1

We will try to extract the key facts needed in the proof of Theorem 1. They are condensed into several lemmas of independent interest. We will also need Theorems 2.1, 2.2, and 2.4. The implication (i) $\Rightarrow$ (v) is established in Lemma 3, and (v) $\Rightarrow$ (iv) is proved in Lemma 2. Since clearly, (iv) $\Rightarrow$ (iii) $\Rightarrow$ (ii) $\Rightarrow$ (i), this completes the proof of Theorem 1.

Throughout this section, we will use the notation

$$g_h(x) = E(f_n(x)) = \int h^{-d} K\left(\frac{x-y}{h}\right) f(y)\, dy. \tag{1}$$

LEMMA 1 (A Multinomial Distribution Inequality). *Let $(X_1, \ldots, X_k)$ be a multinomial $(n, p_1, \ldots, p_k)$ random vector. For $\varepsilon \in (0,1)$ and all k satisfying $k/n \le \varepsilon^2/20$, we have*

$$P\left(\sum_{i=1}^{k} |X_i - E(X_i)| \ge n\varepsilon\right) \le 3e^{-n\varepsilon^2/25}.$$

Proof. The proof is based upon a Poissonization. Let N be a Poisson (n) random variable independent of $U_1, U_2, \ldots$, a sequence of independent $\{1, \ldots, k\}$-valued variables distributed according to $P(U_1 = i) = p_i$, $1 \le i \le k$. Let X_i be the number of occurrences of the value i among $U_1, \ldots, U_n$, and let X_i' be the number of occurrences of the value i among $U_1, \ldots, U_N$. It is clear that $X_1', \ldots, X_k'$ are independent Poisson random variables with means $np_1, \ldots, np_k$, and that $X_1, \ldots, X_k$ is a multinomial $(n, p_1, \ldots, p_k)$ random vector. We have

$$\sum_{i=1}^{k} \frac{1}{n}|X_i - np_i| \le \sum_{i=1}^{k} \frac{1}{n}|X_i - X_i'| + \sum_{i=1}^{k} \frac{1}{n}|X_i' - np_i|. \tag{2}$$

Now, when U is Poisson (λ), then for $t > 0$, $E(e^{t|U-\lambda|}) \le E(e^{t(U-\lambda)} +$

there exist functions $h_0(\varepsilon)$ and $c_0(\varepsilon)$ such that for

$$\left(\frac{c_0(\varepsilon)}{n}\right)^{1/d} < h < h_0(\varepsilon),$$

this exponential bound is valid. Thus, for a given ε, h may remain constant and the exponential inequality remains valid nevertheless.

2. PROOF OF THEOREM 1

We will try to extract the key facts needed in the proof of Theorem 1. They are condensed into several lemmas of independent interest. We will also need Theorems 2.1, 2.2, and 2.4. The implication (i) $\Rightarrow$ (v) is established in Lemma 3, and (v) $\Rightarrow$ (iv) is proved in Lemma 2. Since clearly, (iv) $\Rightarrow$ (iii) $\Rightarrow$ (ii) $\Rightarrow$ (i), this completes the proof of Theorem 1.

Throughout this section, we will use the notation

$$g_h(x) = E(f_n(x)) = \int h^{-d} K\left(\frac{x-y}{h}\right) f(y)\, dy. \tag{1}$$

LEMMA 1 (A Multinomial Distribution Inequality). *Let $(X_1, \ldots, X_k)$ be a multinomial $(n, p_1, \ldots, p_k)$ random vector. For $\varepsilon \in (0,1)$ and all k satisfying $k/n \le \varepsilon^2/20$, we have*

$$P\left(\sum_{i=1}^{k} |X_i - E(X_i)| \ge n\varepsilon\right) \le 3e^{-n\varepsilon^2/25}.$$

Proof. The proof is based upon a Poissonization. Let N be a Poisson (n) random variable independent of $U_1, U_2, \ldots$, a sequence of independent $\{1, \ldots, k\}$-valued variables distributed according to $P(U_1 = i) = p_i$, $1 \le i \le k$. Let X_i be the number of occurrences of the value i among $U_1, \ldots, U_n$, and let X_i' be the number of occurrences of the value i among $U_1, \ldots, U_N$. It is clear that $X_1', \ldots, X_k'$ are independent Poisson random variables with means $np_1, \ldots, np_k$, and that $X_1, \ldots, X_k$ is a multinomial $(n, p_1, \ldots, p_k)$ random vector. We have

$$\sum_{i=1}^{k} \frac{1}{n} |X_i - np_i| \le \sum_{i=1}^{k} \frac{1}{n} |X_i - X_i'| + \sum_{i=1}^{k} \frac{1}{n} |X_i' - np_i|. \tag{2}$$

Now, when U is Poisson (λ), then for $t > 0$, $E(e^{t|U-\lambda|}) \le E(e^{t(U-\lambda)} +$

$e^{t(\lambda - U)}) = e^{\lambda(e^t - 1) - t\lambda} + e^{\lambda(e^{-t} - 1) + t\lambda} \le 2e^{\lambda(e^t - 1 - t)}$, because $e^{-t} + t \le e^t - t$. Thus,

$$P(|U - \lambda| \ge \lambda\varepsilon) \le E(e^{t|U-\lambda| - t\lambda\varepsilon}) \le 2e^{-t\lambda\varepsilon}e^{\lambda(e^t - 1 - t)}$$

$$= 2e^{\lambda(\varepsilon - (1+\varepsilon)\ln(1+\varepsilon))} \le 2e^{-\lambda\varepsilon^2/2(1+\varepsilon)} \le 2e^{-\lambda\varepsilon^2/4}, \quad (3)$$

where we took $t = \ln(1 + \varepsilon)$. By a repetition of the previous argument,

$$P\left(\sum_{i=1}^{k} \frac{1}{n}|X_i - np_i| \ge \varepsilon\right)$$

$$\le P\left(|N - n| \ge n\frac{2\varepsilon}{5}\right) + P\left(\sum_{i=1}^{k} |X_i' - np_i| \ge n\frac{3\varepsilon}{5}\right)$$

$$\le 2e^{-n(2\varepsilon/5)^2/4} + e^{-tn(3\varepsilon/5)}\prod_{i=1}^{k}\left(2e^{np_i(e^t - 1 - t)}\right) \quad \text{(by (3))}$$

$$\le 2e^{-n\varepsilon^2/25} + 2^k e^{n(e^t - 1 - t - 3\varepsilon t/5)}$$

$$\le 2e^{-n\varepsilon^2/25} + e^{k - n(3\varepsilon/5)^2/4} \quad (\text{for } t = \ln(1 + 3\varepsilon/5))$$

$$\le 3e^{-n\varepsilon^2/25} \quad (\text{when } k \le n\varepsilon^2/20). \quad (4)$$

LEMMA 2. *For any density f on R^d, and any absolutely integrable function K with $\int K(x)\,dx = 1$,* (iv) *holds whenever*

$$\lim_{n\to\infty} h = 0 \quad \textit{and} \quad \lim_{n\to\infty} nh^d = \infty.$$

Proof. Let g_h be defined as in (1). By Theorem 2.1, it suffices to show that $\int |f_n(x) - g_h(x)|\,dx \to 0$ exponentially. Let μ_n be the empirical probability measure for $X_1, \ldots, X_n$, and note that

$$f_n(x) = h^{-d}\int K\left(\frac{x - y}{h}\right)\mu_n(dy).$$

For given $\varepsilon > 0$, find finite constants $M, L, N, a_1, \ldots, a_N$ and disjoint finite rectangles $A_1, \ldots, A_N$ in R^d such that the function

$$K^*(x) = \sum_{i=1}^{N} a_i I_{A_i}(x)$$

satisfies: $|K^*| \le M$, $K^* = 0$ outside $[-L, L]^d$, and $\int |K(x) - K^*(x)|\,dx < \varepsilon$. Define g_h^* and f_n^* as g_h and f_n with K^* instead of K. Then

$$\begin{aligned}\int |f_n(x) - g_h(x)|\,dx &\le \int |f_n(x) - f_n^*(x)|\,dx + \int |f_n^*(x) - g_h^*(x)|\,dx \\ &\quad + \int |g_h^*(x) - g_h(x)|\,dx \\ &\le \int h^{-d} \int \left|K^*\left(\frac{x-y}{h}\right) - K\left(\frac{x-y}{h}\right)\right| f(y)\,dy\,dx \\ &\quad + \int h^{-d} \int \left|K^*\left(\frac{x-y}{h}\right) - K\left(\frac{x-y}{h}\right)\right| \mu_n(dy)\,dx \\ &\quad + \int |f_n^*(x) - g_h^*(x)|\,dx \\ &\le 2\varepsilon + \int |f_n^*(x) - g_h^*(x)|\,dx\end{aligned}$$

by a double change of integral. But if μ is the probability measure for f,

$$\begin{aligned}\int |f_n^*(x) - g_h^*(x)|\,dx &\le \sum_{i=1}^N |a_i| \int \left| h^{-d} \int_{x+hA_i} f(y)\,dy - h^{-d} \int_{x+hA_i} \mu_n(dy) \right| dx \\ &\le Mh^{-d} \sum_{i=1}^N \int |\mu(x + hA_i) - \mu_n(x + hA_i)|\,dx.\end{aligned}$$

Lemma 2 follows if we can show that for all finite rectangles A of R^d,

$$h^{-d} \int |\mu(x + hA) - \mu_n(x + hA)|\,dx \to 0 \quad \text{exponentially as } n \to \infty.$$

Choose an A, and let $\varepsilon > 0$ be arbitrary. Consider the partition of R^d into sets B that are d-fold products of intervals of the form $[(i-1)h/N, ih/N)$, where i is an integer, and N is a new constant to be chosen later. Call the partition Ψ. Let

$$A = \prod_{i=1}^d [x_i, x_i + a_i), \quad \min_i a_i \ge \frac{2}{N}$$

and

$$A^* = \prod_{i=1}^{d} [x_i + 1/N, x_i + a_i - 1/N).$$

Define

$$C_x = x + hA - \bigcup_{\substack{B \in \Psi \\ B \subseteq x + hA}} B \subseteq x + h(A - A^*) = C_x^*.$$

Clearly,

$$\int |\mu(x + hA) - \mu_n(x + hA)|\, dx$$

$$\leq \int \sum_{\substack{B \in \Psi \\ B \subseteq x + hA}} |\mu(B) - \mu_n(B)|\, dx + \int (\mu(C_x^*) + \mu_n(C_x^*))\, dx. \tag{5}$$

The last term in (5) equals

$$2\lambda(h(A - A^*)) = 2h^d \lambda(A - A^*) = 2h^d \left(\prod_{i=1}^{d} a_i - \prod_{i=1}^{d} \left(a_i - \frac{2}{N} \right) \right)$$

$$= 2h^d \lambda(A) \left(1 - \prod_{i=1}^{d} \left(1 - \frac{2}{Na_i} \right) \right)$$

$$\leq 4h^d \lambda(A) \sum_{i=1}^{d} \frac{a_i^{-1}}{N} \leq \varepsilon h^d$$

by choice of N. We used the fact that for any set C, and any probability measure ν on the Borel sets of R^d, $\int \nu(x + hC)\, dx = \lambda(hC)$. For any finite constant $R > 0$, we can bound the first term in (5) from above by

$$\sum_{\substack{B \in \Psi \\ B \cap S_{0R} \neq \emptyset}} |\mu_n(B) - \mu(B)| \int_{B \subseteq x + hA} dx$$

$$+ \int_{B \subseteq x + hA} dx (\mu_n(S_{0R}^c) - \mu(S_{0R}^c) + 2\mu(S_{0R}^c)). \tag{6}$$

Here $(\cdot)^c$ denotes the complement of a set. Clearly, $h^{-d}\int_{B \subseteq x+hA} dx \leq \lambda(A)$, and $\mu(S_{0R}^c) < \varepsilon$ by our choice of R. Also,

$$P\left(\mu_n(S_{0R}^c) - \mu(S_{0R}^c) > \varepsilon\right) \leq e^{-2n\varepsilon^2}$$

by Hoeffding's inequality for binomial random variables (Hoeffding, 1963). Finally, since the collection of sets $B \in \Psi$ with $B \cap S_{0R} \neq \varnothing$ has at most $(2RN/h + 2)^d = o(n)$ elements, we see that by Lemma 3, for all n large enough,

$$P\left(\sum_{\substack{B \in \Psi \\ B \cap S_{0R} \neq \varnothing}} |\mu_n(B) - \mu(B)| > \varepsilon\right) \leq 3e^{-n\varepsilon^2/25}.$$

Now collect bounds. This concludes the proof of Lemma 2.

LEMMA 3. *Let K and f be densities on R^d. If $J_n \to 0$ in probability as $n \to \infty$, then $\lim_{n\to\infty} h = 0$ and $\lim_{n\to\infty} nh^d = \infty$.*

Proof. Since $J_n \leq 2$ for all n, $J_n \to 0$ in probability if and only if $\lim_{n\to\infty} E(J_n) = 0$. Define g_h as in (1). Then

$$E(J_n) = E\left(\int |f_n(x) - f(x)|\,dx\right) \geq \int |E(f_n(x)) - f(x)|\,dx$$

$$= \int |g_h(x) - f(x)|\,dx.$$

By Theorem 2.4, we conclude that $\lim_{n\to\infty} h = 0$. This will be assumed for the remainder of the proof. For the second part, we note that by Theorem 2.1, $E(\int |f_n(x) - g_h(x)|\,dx) \to 0$. Let M be a large number, and let $K^*(x)$ be defined as $K(x)I_{[K(x) \leq M]}$. Define f_n^* and g_h^* as f_n, g_h with K^* instead of K. By Theorem 2.1,

$$\int |f_n(x) - g_h(x)|\,dx \geq \int |f_n^*(x) - g_h^*(x)|\,dx - \int |f_n(x) - f_n^*(x)|\,dx$$

$$- \int |g_h(x) - g_h^*(x)|\,dx$$

$$= \int |f_n^*(x) - g_h^*(x)|\,dx - 2\int |K(x) - K^*(x)|\,dx. \tag{7}$$

Let us introduce some more notation: L is another large number, A is the

event that no $X_i, 1 \le i \le n$, belongs to $S_{x\,hL}$, $K' = K^* I_{S_{0\,L}}$, $K'' = K^* - K'$, and f_n' and f_n'' are defined as f_n after replacement of K by K' and K'' in the definition. Clearly,

$$\begin{aligned}\int E(|f_n^*(x) - g_h^*(x)|\,dx) &\ge \int E(|f_n^*(x) - g_h^*(x)|I_A)\,dx \\ &\ge \int g_h^*(x)P(A)\,dx - \int E(f_n''(x)I_A)\,dx \\ &= U_n - V_n. \end{aligned} \tag{8}$$

We will need the following facts, all corollaries of Theorems 2.2 and 2.3: for bounded K^* with compact support, $g_h^*(x) \to f(x)\int K^*(x)\,dx$, almost all x, and $\mu(S_{y+hz\ hL})/\lambda(S_{y+hz\ hL}) \to f(y)$ for all $z \in R^d$ and almost all $y \in R^d$. Let C be the volume of $S_{0\,1}$, and assume that $\lim_{n\to\infty} nh^d = s \in [0, \infty)$. By Fatou's Lemma, we have

$$\begin{aligned}\liminf_{n\to\infty} U_n &\ge \int \liminf_{n\to\infty} g_h'(x) \liminf_{n\to\infty} P(A)\,dx \\ &= \int f(x) \liminf_{n\to\infty} (1 - \mu(S_{x\,hL}))^n\,dx \int K'(z)\,dz \\ &\ge \int f(x)\exp\left(-\limsup_{n\to\infty}\left(\frac{n\mu(S_{x\,hL})}{1-\mu(S_{x\,hL})}\right)\right)dx \int K'(z)\,dz \\ &= \int f(x)\exp(-sCL^d f(x))\,dx \int_{S_{0\,L}} K^*(z)\,dz. \end{aligned} \tag{9}$$

Also,

$$\begin{aligned} V_n &\le \int E\left(\frac{1}{n}\sum_{i=1}^{n} h^{-d}K''\left(\frac{x - X_i}{h}\right)I_A\right)dx \\ &= \int\int h^{-d}K''\left(\frac{x-y}{h}\right)I_{y \notin S_{x\,hL}} f(y)\,dy\,(1 - \mu(S_{x\,hL}))^{n-1}\,dx \\ &= \int f(y)\int_{x \notin S_{y\,hL}} h^{-d}K''((x-y)/h)(1 - \mu(S_{x\,hL}))^{n-1}\,dx\,dy \\ &\le \int f(y)\int_{z \notin S_{0\,L}} K''(z)\exp(-(n-1)\mu(S_{y+hz\,hL}))\,dz\,dy. \end{aligned} \tag{10}$$

The integrand of the inner integral of (10) is bounded by an integrable

function, K''. Thus, by the Lebesgue dominated convergence theorem and an earlier remark, we can conclude that

$$\limsup_{n\to\infty} V_n \le \int f(y)\int_{z\notin S_{0,L}} K^*(z)\exp(-sCL^d f(y))\,dz\,dy$$

$$= \int f(y)\exp(-sCL^d f(y))\,dy \int_{z\notin S_{0,L}} K^*(z)\,dz. \qquad (11)$$

Combining (7), (8), (9), and (11) gives

$$\liminf_{n\to\infty} \int E(|f_n(x) - g_h(x)|)\,dx + 2\int |K(x) - K^*(x)|\,dx$$

$$\ge \int f(x)\exp(-sCL^d f(x))\,dx\left(2\int_{S_{0,L}} K^*(z)\,dz - 1\right). \qquad (12)$$

Since M was arbitrary, we have

$$\liminf_{n\to\infty} \int E(|f_n(x) - g_h(x)|)\,dx \ge \int f e^{-sCL^d f}\left(2\int_{S_{0,L}} K - 1\right).$$

Now, choose L finite but large enough so that $\int_{S_{0,L}} K > \frac{1}{2}$. Then, in order for the right-hand side of the last inequality to be 0, we must have $s = \infty$, and this is a contradiction. Thus, no subsequence of nh^d can tend to a finite limit s, and therefore, we must have $\lim_{n\to\infty} nh^d = \infty$.

3. HISTOGRAM ESTIMATE

The *histogram estimate* is defined by a sequence of partitions $\mathscr{P}_n = \{A_{nj}, j = 1, 2, \ldots\}, n \ge 1$, where all A_{nj}'s are Borel sets with finite nonzero Lebesgue measure. We assume that the sequence of partitions is rich enough such that the class of Borel sets ($\mathscr{B}$) is equal to

$$\bigcap_{n=1}^{\infty} \sigma\left(\bigcup_{m=n}^{\infty} \mathscr{P}_m\right), \qquad (13)$$

where we use the symbol σ for the σ-algebra generated by a class of sets. The histogram estimate is defined by

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{I_{[X_i \in A_{nj}]}}{\lambda(A_{nj})}, \qquad x \in A_{nj},$$

and its expected value is

$$g_n(x) = E(f_n(x)) = \int_{A_{nj}} \frac{f}{\lambda(A_{nj})}, \qquad x \in A_{nj}.$$

Abou-Jaoude (1976a, c) prove the following.

THEOREM 2. *Assume that the sequence of partitions $\mathscr{P}_n$ satisfies* (13). *Then the following conditions are equivalent:*

(i) $J_n \to 0$ *in probability as* $n \to \infty$, *all* f.

(ii) $J_n \to 0$ *almost surely as* $n \to \infty$, *all* f.

(iii) $J_n \to 0$ *exponentially as* $n \to \infty$ (*see Theorem* 1), *all* f (*as in Theorem* 1, *the exponent can be taken independently of* f *and of the partition*).

(iv) *For all* $A \in \mathscr{B}$ *with* $0 < \lambda(A) < \infty$, *and all* $\varepsilon > 0$ *there exists* n_0 *such that for all* $n \geq n_0$ *we can find* $A_n \in \sigma(\mathscr{P}_n)$ *with* $\lambda(A \Delta A_n) < \varepsilon$; (14)

and

$$\sup_{\substack{M>0 \\ \text{all sets } C \text{ of} \\ \text{finite Lebesgue} \\ \text{measure}}} \limsup_{n\to\infty} \lambda\left(\bigcup_{j:\lambda(A_{nj}\cap C)\leq M/n} A_{nj} \cap C \right) = 0. \quad (15)$$

Our proof differs from Abou-Jaoude's only in the details. For example, the powerful Lemma 1 provides us with a shortcut. Conditions (14) and (15) are sometimes easy to check. Consider, for example, the *cubic histogram estimate* where each A_{nj} is of the type $\prod_{i=1}^{d} [a_i k_i h, a_i(k_i + 1)h)$ where the k_i's are integers, h is a smoothing factor as for the kernel estimate, and the a_i's are positive constants. In Theorem 2.5, we have shown that (14) holds for this estimate if and only if

$$\lim_{n\to\infty} h = 0.$$

Furthermore, it is easy to see that (15) holds if and only if

$$\lim_{n\to\infty} nh^d = \infty.$$

We should point out that in another paper, Abou-Jaoude (1976b) has given similar necessary and sufficient conditions for weak convergence in L_1

of histogram estimates in R^1 in which the partitions depend upon the order statistics of $X_1, \ldots, X_n$. See Theorem 7.3.

4. PROOF OF THEOREM 2

We always have $E(\int |f_n - f|) \geq \int |g_n - f|$. Now, by the boundedness of J_n, convergence in the mean and in probability to 0 are equivalent. Thus, (i) $\Rightarrow$ (14), by Theorem 2.5. Since obviously (iii) $\Rightarrow$ (ii) $\Rightarrow$ (i), we are just left with the proofs of (iv) $\Rightarrow$ (iii) and (i) $\Rightarrow$ (15). Again, this will be done in two separate lemmas.

LEMMA 4. (iv) $\Rightarrow$ (iii).

Proof. We know that $\int |g_n - f| \to 0$ (Theorem 2.5). Thus, it suffices to show that $\int |f_n - g_n| \to 0$ exponentially. Let μ_n be the empirical measure for $X_1, \ldots, X_n$, and let μ be the probability measure defined by f on the Borel sets of R^d. We have

$$\int |f_n - g_n| = \sum_j |\mu_n(A_{nj}) - \mu(A_{nj})|.$$

Divide the positive integers into two sets, H_n and its complement H_n^c, where H_n collects all integers j for which $\lambda(A_{nj}C) > M/n$. We have

$$\begin{aligned} \int |f_n - g_n| &\leq \sum_{j \in H_n} |\mu_n(A_{nj}) - \mu(A_{nj})| + \sum_{j \in H_n^c} \left(\mu_n(A_{nj}) + \mu(A_{nj})\right) \\ &\leq \sum_{j \in H_n} |\mu_n(A_{nj}) - \mu(A_{nj})| + 2\mu(A_n) + |\mu_n(A_n) - \mu(A_n)|, \end{aligned} \tag{16}$$

where $A_n = \bigcup_{j \in H_n^c} A_{nj}$.

Since $\lambda(A_{nj} \cap C) > M/n$ for $j \in H_n$, H_n cannot have more than $1 + n\lambda(C)/M$ members. Also, $\{n\mu_n(A_n); n\mu_n(A_{nj}C), j \in H_n\}$ is multinomially distributed. Therefore, by Lemma 1, when

$$\frac{\lambda(C)}{M} + \frac{2}{n} \leq \frac{\varepsilon^2}{20}, \tag{17}$$

we have

$$P\left(\sum_{j \in H_n} |\mu_n(A_{nj}) - \mu(A_{nj})| + |\mu_n(A_n) - \mu(A_n)| > \varepsilon \right) \leq 3 \exp\left(-\frac{n\varepsilon^2}{25} \right).$$

We can make $\mu(A_n)$ as small as desired by choice of C. Indeed,

$$\mu(A_n) = \mu(A_n \cap C) + \mu(A_n \cap C^c) \le o(1) + \mu(C^c), \tag{18}$$

where the "$o(1)$" part follows from (15) (which states that $\lambda(A_n \cap C) = o(1)$) and the fact that μ is absolutely continuous with respect to λ.

Thus, for given $\varepsilon > 0$, choose C such that (18) $\le \varepsilon + o(1)$, and then choose M so that (17) holds for all n large enough. Combining all the inequalities in (16) gives

$$P\left(\int |f_n - g_n| > 4\varepsilon\right) \le 3\exp\left(-\frac{n\varepsilon^2}{25}\right), \quad \text{all } n \text{ large enough.}$$

This concludes the proof of Lemma 4.

LEMMA 5. $(i) \Rightarrow (15)$.

Proof. We keep the notation of Lemma 4. In particular, $M > 0$ is a constant, and C is a set of finite Lebesgue measure. Assume that $\lambda(C) > 0$, and define $f = I_C/\lambda(C)$, and

$$Z_n = \sum_j \lambda(A_{nj} \cap C) \frac{I_{[A_{nj} \cap C \text{ does not capture any } X_i]}}{\lambda(C)}.$$

We have

$$\int |f_n - g_n| = \sum_j |\mu_n(A_{nj}) - \mu(A_{nj})| = \sum_j \left|\mu_n(A_{nj}) - \frac{\lambda(A_{nj} \cap C)}{\lambda(C)}\right| \ge Z_n.$$

Since $0 \le Z_n \le 1$, (i) implies that $E(Z_n) \to 0$. Now,

$$\begin{aligned} E(Z_n) &= \sum_j \frac{\lambda(A_{nj} \cap C)}{\lambda(C)} \left(1 - \frac{\lambda(A_{nj} \cap C)}{\lambda(C)}\right)^n \\ &\ge \sum_{j \in H_n^c} \frac{\lambda(A_{nj} \cap C)}{\lambda(C)} \left(1 - \frac{M}{n\lambda(C)}\right)^n \\ &\ge \frac{\lambda(A_n \cap C)}{\lambda(C)} \exp\left(-\frac{M/\lambda(C)}{1 - M/n\lambda(C)}\right) \\ &\sim \frac{\lambda(A_n \cap C)}{\lambda(C)} \exp\left(-\frac{M}{\lambda(C)}\right). \end{aligned}$$

t this implies that for every C with $\lambda(C) > 0, \lim_{n\to\infty} \lambda(A_n \cap C) = 0$. so, when $\lambda(C) = 0$, it is clear that $\lambda(A_n \cap C) = 0$ for all n. Thus, (15) is isfied.

RELATIVE STABILITY

compare different density estimates, it is inconvenient to work with the ıdom variable J_n. One could use the quantiles or the moments of J_n. We ll use $E(J_n)$ throughout the rest of this book. This would simply be a poor ɔice if J_n were not close to $E(J_n)$ in some sense. In fact, we would like our imates to be *relatively stable* (in probability, almost surely), that is, we ıuld like

$$\frac{J_n}{E(J_n)} \to 1 \quad \text{(in probability, almost surely)}. \tag{19}$$

ɔte that a sequence of random variables J_n is usually said to be relatively ıble when there exists a sequence of real numbers a_n such that J_n/a_n ıds to one in some stochastic mode. Our definition differs slightly because : force $a_n = E(J_n)$. Proving (19) however is virtually as difficult as termining the limit law of J_n. Fortunately, it is much easier to prove that e *variation* of an estimate, $\int |f_n - E(f_n)|$, is relatively stable. Via Lemmas and 7, this yields statements that come close to (19) for J_n.

EMMA 6. *For any density f on R^d, and any density estimate f_n,*

$$\text{Max}\left(\int |f - E(f_n)|, \frac{1}{2}\int |f_n - E(f_n)|\right) \le \int |f_n - f|$$

$$\le \int |f_n - E(f_n)| + \int |f - E(f_n)|.$$

roof. For the lower bound, note that, by Jensen's inequality, $\int |f_n - f| \ge \int |E(f_n) - f|$. Also, by the triangle inequality, $\int |f_n - f| \ge \int |f_n - E(f_n)| - \int |E(f_n) - f|$. Summing both inequalities gives the desired result.

EMMA 7. *If the variation of a density estimate is relatively stable in robability, that is,*

$$\frac{\int |f_n - E(f_n)|}{E\left(\int |f_n - E(f_n)|\right)} \to 1 \quad \textit{in probability},$$

then $P(J_n/E(J_n) \notin (\frac{1}{4} - \varepsilon, 3 + \varepsilon)) \to 0$ *as* $n \to \infty$, *all* $\varepsilon > 0$. *If the variation is relatively stable almost surely, then* $P(J_n/E(J_n) \notin (\frac{1}{4} - \varepsilon, 3 + \varepsilon)$ *i.o.*$) = 0$, *all* $\varepsilon > 0$.

Proof. We use Lemma 6 twice. First, for the upper bound, note that

$$\frac{J_n}{E(J_n)} \le \frac{\int|f_n - E(f_n)| + \int|f - E(f_n)|}{E\left(\max\left(\int|f - E(f_n)|, \frac{1}{2}\int|f_n - E(f_n)|\right)\right)}$$

$$\le \frac{\int|f_n - E(f_n)| + \int|f - E(f_n)|}{\max\left(\int|f - E(f_n)|, \frac{1}{2}E\left(\int|f_n - E(f_n)|\right)\right)}$$

$$\le 2\frac{\int|f_n - E(f_n)|}{E\left(\int|f_n - E(f_n)|\right)} + 1,$$

where we used Jensen's inequality. For the lower bound, we let A be the (deterministic) event $[\int|f - E(f_n)| \ge \frac{1}{2}E(\int|f_n - E(f_n)|)]$, and note that

$$\frac{J_n}{E(J_n)} \ge \frac{\max\left(\int|f - E(f_n)|, \frac{1}{2}\int|f_n - E(f_n)|\right)}{E\left(\int|f_n - E(f_n)|\right) + \int|f - E(f_n)|}$$

$$\ge \frac{1}{2}I_A + \frac{1}{4}I_{A^c}\frac{\int|f_n - E(f_n)|}{E\left(\int|f_n - E(f_n)|\right)}$$

$$\ge \min\left(\frac{1}{2}, \frac{1}{4}\frac{\int|f_n - E(f_n)|}{E\left(\int|f_n - E(f_n)|\right)}\right).$$

Lemma 7 follows directly from these inequalities.

Roughly speaking, if the variation of an estimate is relatively stable, then $J_n/E(J_n)$ remains with high probability in $[\frac{1}{4}, 3]$. This would indicate that $E(J_n)$ is a fairly good yardstick for comparing density estimates. (This is precisely what we will do in Chapters 4, 5, 7, 8, and 9.)

In the remainder of this section, we follow Abou-Jaoude (1977), who showed that the histogram estimate and the kernel estimate with a uniform kernel $K(x) = I_{[-1/2,1/2]^d}$ have variations that are relatively stable in probability for *all* f.

To do this, we will need a few inequalities for the binomial distribution.

LEMMA 8 (Inequality for the Absolute Deviation of a Binomial Random Variable). *Let X be a binomial (n, p) random variable with $p \le \frac{1}{2}$. Then,*

$$E\left(\left(p - \frac{X}{n}\right)_+\right) \ge \begin{cases} p/e^2 & \text{if } p < 1/n; \\ c\sqrt{p/n} & \text{if } p \ge 1/n; \end{cases}$$

where $c = (\sqrt{4\pi}\, e^{13/6})^{-1}$ is a universal constant. Also,

$$E\left(\left(p - \frac{X}{n}\right)_+\right) \le \sqrt{\frac{p}{n}}\,.$$

Proof. Let $m = \underline{np}$ (the largest integer contained in np). We have, by elementary computations, for $n \ge 2$,

$$E((np - X)_+) = \sum_{i=0}^{m} (np - i)\binom{n}{i} p^i (1-p)^{n-i}$$

$$= np\binom{n-1}{m} p^m (1-p)^{n-m}.$$

If $p < 1/n$, and thus $m = 0$, this is equal to $np(1-p)^n \ge npe^{-np/(1-p)} \ge npe^{-2}$. If $p \ge 1/n$, and thus $m > 0$, we obtain, by Stirling's formula,

$$E((np - X)_+) = (2\pi)^{-1/2} p \sqrt{\frac{n-m}{m/n}} \left(\frac{np}{m}\right)^m \left(\frac{n(1-p)}{n-m}\right)^{n-m} g(n,m),$$

where $g(n,m) = \exp(u/12n - v/12m - w/12(n-m)) \ge \exp(-\frac{1}{6})$ (here u, v, w are numbers in $[0,1]$). Also, $n - m \ge n/2$, and $m/n \le p$. Further-

more, since $m = np - z$, $z \in [0, 1]$,

$$\left(\frac{np}{m}\right)^m \left(\frac{n(1-p)}{n-m}\right)^{n-m} = \left(1 - \frac{z}{np}\right)^{-(np-z)} \left(1 + \frac{z}{n(1-p)}\right)^{-(n-np+z)}$$

$$\geq \exp\left(+z - \frac{z^2}{np} - z - \frac{z^2}{n-np}\right)$$

$$\geq \exp\left(-\frac{1}{np} - \frac{1}{n-np}\right) \geq \exp(-2).$$

Combining these estimates gives our result for $n \geq 2$.

For $n = 1$, note that $E((p - X/n)_+) = p(1-p)$, and, thus, our inequality follows for all n. The upper bound is obtained simply by using the Cauchy–Schwarz inequality and noting that $E((p - X/n)^2) = p(1-p)/n \leq p/n$.

LEMMA 9 (Geffroy. See Abou-Jaoude, 1977, pp. 52–53). *Let p_1, p_2, p_3 be a probability vector, and let X_1, X_2, X_3 be multinomial (n, p_1, p_2, p_3) random vector. Then*

$$E\left(\left(p_1 - \frac{X_1}{n}\right)_+ \left(p_2 - \frac{X_2}{n}\right)_+\right) \leq E\left(\left(p_1 - \frac{X_1}{n}\right)_+\right) E\left(\left(p_2 - \frac{X_2}{n}\right)_+\right).$$

Proof. Assume that (20) is valid:

$$E\left(\left(p_1 - \frac{X_1}{n}\right)_+ \middle| X_2 = n_2\right) \text{ is increasing and convex in } n_2. \tag{20}$$

Then,

$$E\left(\left(p_1 - \frac{X_1}{n}\right)_+\right) = \sum_{n_2=0}^{n} P(X_2 = n_2) E\left(\left(p_1 - \frac{X_1}{n}\right)_+ \middle| X_2 = n_2\right)$$

$$\geq E\left(\left(p_1 - \frac{X_1}{n}\right)_+ \middle| X_2 = np_2\right) \quad \text{(Jensen's inequality)}$$

$$\geq E\left(\left(p_1 - \frac{X_1}{n}\right)_+ \middle| X_2 = n_2\right), \quad \text{all } n_2 \leq np_2.$$

Thus,

$$E\left(\left(p_1 - \frac{X_1}{n}\right)_+\left(p_2 - \frac{X_2}{n}\right)_+\right) = \sum_{n_2=0}^{np_2}\left(p_2 - \frac{1}{n}n_2\right)P(X_2 = n_2)$$

$$\times E\left(\left(p_1 - \frac{X_1}{n}\right)_+ \middle| X_2 = n_2\right)$$

$$\leq E\left(\left(p_1 - \frac{X_1}{n}\right)_+\right)E\left(\left(p_2 - \frac{X_2}{n}\right)_+\right),$$

which was to be shown.

Let us prove (20) now. We must show that $\psi(m) = \phi(m+1) - \phi(m)$ is positive and increasing in m, where ϕ is the function defined in (20). Let Y_m be the random variable X_1 given that $X_2 = m$. Obviously, we have the following embedding: $Y_m = Y_{m+1} + Z$ where Z is a Bernoulli random variable with parameter $p_1/(1-p_2)$, and Z is independent of Y_{m+1}. Thus,

$$\psi(m) = E\left(\left(p_1 - \frac{Y_{m+1}}{n}\right)_+ - \left(p_1 - \frac{Y_{m+1}}{n} - Z\right)_+\right) = E(U).$$

But U takes the following values:

$$U = \begin{cases} 0 & \text{if } Z = 0, \text{ or if } Z = 1,\ p_1 - (1/n)Y_{m+1} \leq 0; \\ p_1 - (1/n)Y_{m+1} & \text{if } Z = 1,\ 0 < p_1 - (1/n)Y_{m+1} \leq 1/n; \\ 1/n & \text{if } Z = 1,\ 1/n < p_1 - (1/n)Y_{m+1}. \end{cases}$$

If $z = np_1 - \underline{np_1}$, then

$$\psi(m) = P(Z=1)\left(\frac{z}{n}P(Y_{m+1} = \underline{np_1}) + \frac{1}{n}P(Y_{m+1} \leq np_1 - 1)\right)$$

$$= \frac{p_1}{n(1-p_2)}\left(zP(Y_{m+1} \leq np_1) + (1-z)P(Y_{m+1} \leq np_1 - 1)\right).$$

This expression is positive. By our embedding, we also note that it increases with m. This concludes the proof of Lemma 9.

LEMMA 10. *Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of nonnegative random variables with $E(Z_n) \neq 0$, for all n, and $E(Z_n^2) < \infty$. Then $Z_n/E(Z_n) \to 1$*

in probability whenever

$$\lim_{n\to\infty} \frac{E(Z_n^2)}{(E(Z_n))^2} = 1.$$

Proof. By Chebyshev's inequality, for all $\epsilon > 0$,

$$P\left(\left|\frac{Z_n}{E(Z_n)}\right| > 1+\epsilon\right) \le \frac{\mathrm{Var}(Z_n)}{\epsilon^2 E^2(Z_n)} = o(1).$$

THEOREM 3 (Abou-Jaoude, 1977). *Assume that for the histogram estimate of Section 3, there exists a constant $\eta > 0$, such that for all $\epsilon > 0$, and some n_0,*

$$A_n(\epsilon) = \sum_{\mu(A_{nj}) \le \epsilon} \mu(A_{nj}) \ge \eta, \qquad n \ge n_0.$$

(This is satisfied for the cubic histogram estimate when $h \to 0$.) Then the variation of the histogram estimate is relatively stable in probability:

$$\frac{\int |f_n - E(f_n)|}{E\left(\int |f_n - E(f_n)|\right)} \to 1 \quad \textit{in probability.}$$

Proof. We note that

$$\int |f_n - E(f_n)| = 2Z_n,$$

where

$$Z_n = \sum_j \left(\mu(A_{nj}) - \mu_n(A_{nj})\right)_+.$$

In view of Lemma 10 and the obvious inequality $E^2(Z_n) \le E(Z_n^2)$, we need only show that

$$\limsup_{n\to\infty} \frac{E(Z_n^2)}{E^2(Z_n)} \le 1. \tag{21}$$

But, using Lemma 9, we have

$$E(Z_n^2) = \sum_j E\Big(\big(\mu(A_{nj}) - \mu_n(A_{nj})\big)_+^2\Big)$$

$$+ \sum_{i \neq j} E\Big(\big(\mu(A_{nj}) - \mu_n(A_{nj})\big)_+ \big(\mu(A_{ni}) - \mu_n(A_{ni})\big)_+\Big)$$

$$\leq \frac{1}{n} \sum_j \mu(A_{nj})\big(1 - \mu(A_{nj})\big)$$

$$+ \sum_{i \neq j} E\Big(\big(\mu(A_{nj}) - \mu_n(A_{nj})\big)_+\Big) E\big(\big(\mu(A_{ni}) - \mu_n(A_{ni})\big)_+\big)$$

$$\leq \frac{1}{n} + E^2(Z_n). \tag{22}$$

But (21) follows from (22) and $\sqrt{n}\,E(Z_n) \to \infty$, which we shall now show. By Lemma 8, and using the constant c from that Lemma,

$$\sqrt{n}\,E(Z_n) \geq c \sum_{\varepsilon \geq \mu(A_{nj}) \geq 1/n} \sqrt{\mu(A_{nj})} + e^{-2}\sqrt{n} \sum_{\mu(A_{nj}) < 1/n} \mu(A_{nj})$$

$$\geq \eta \min\left(\frac{c}{\sqrt{\varepsilon}}, \frac{\sqrt{n}}{e^2}\right), \qquad n \geq n_0.$$

Now, Theorem 3 follows from the arbitrariness of ε.

REMARK. For the cubic histogram estimate with smoothing factor $h = h_n \to 0$, we know that $\sup \mu(A_{nj}) \to 0$ (by the absolute continuity of μ with respect to Lebesgue measure), and thus that for any $\varepsilon > 0$,

$$\lim_{n \to \infty} A_n(\varepsilon) = 1.$$

Thus, the condition of Theorem 3 is satisfied.

THEOREM 4 (Abou-Jaoude, 1977). *Consider the kernel estimate with kernel* $K(x) = I_{[-1/2,1/2]^d}$ *and smoothing factor* $h \to 0$. *Then the variation is relatively stable in probability.*

Proof. We will use C for $[-\frac{1}{2}, \frac{1}{2}]^d$, $p(x)$ for $\mu(x + hC)$, and $p_n(x)$ for $\mu_n(x + hC)$. Recall that $\sup_x p(x) \to 0$ as $n \to \infty$, by the absolute continuity of μ with respect to Lebesgue measure. Arguing as in the proof of

Theorem 3, we note that the variation is $2Z_n/h^d$, where

$$Z_n = \int (p - p_n)_+.$$

Again, it suffices to establish (21). Now, let D be the collection of all x, y in R^{2d} for which the L_∞ distance $\|x - y\| \le h$, and let D^c be its complement. Note that by Lemma 9, applied to D^c,

$$E(Z_n^2) = \int\int E\big((p(x) - p_n(x))_+ (p(y) - p_n(y))_+\big)\, dx\, dy$$

$$\le \int_D E\big((p(x) - p_n(x))_+ (p(y) - p_n(y))_+\big)\, dx\, dy + E^2(Z_n).$$

$$= a_n + E^2(Z_n).$$

We are done if we can show that $E(Z_n)/\sqrt{a_n} \to \infty$. To do so, we must obtain good upper bounds for a_n.

Consider now for fixed x, y the sets $A = (x + hC) - (y + hC)$, $A' = (x + hC) \cap (y + hC)$ and $A'' = (y + hC) - (x + hC)$. Then,

$$(p(x) - p_n(x))_+ (p(y) - p_n(y))_+$$

$$= (\mu(A + A') - \mu_n(A + A'))_+ (\mu(A' + A'') - \mu_n(A' + A''))_+$$

$$\le \big((\mu(A) - \mu_n(A))_+ + (\mu(A') - \mu_n(A'))_+\big)$$

$$\times \big((\mu(A') - \mu_n(A'))_+ + (\mu(A'') - \mu_n(A''))_+\big)$$

$$\le (\mu(A) - \mu_n(A))_+^2 + 2(\mu(A') - \mu_n(A'))_+^2 + (\mu(A'') - \mu_n(A''))_+^2,$$

and the expected value of this does not exceed

$$\frac{1}{n}(\mu(A) + 2\mu(A') + \mu(A'')) \le \frac{1}{n}(\mu(x + hC) + \mu(y + hC)).$$

Thus,

$$a_n \le \int_D \frac{1}{n}(\mu(x + hC) + \mu(y + hC))\,dx\,dy$$

$$\le \frac{2}{n}(2h)^d \int \mu(x + hC)\,dx$$

$$= \frac{2}{n}2^d h^{2d}.$$

But by Lemma 8, if $\sup p(x) \le \frac{1}{2}$,

$$E(Z_n) \ge \int_{1/2 \ge p(x) > 1/n} \frac{c}{\sqrt{n}} \frac{p(x)}{\sqrt{\sup p(x)}}\,dx + \int_{p(x) \le 1/n} e^{-2} p(x)\,dx$$

$$\ge \frac{h^d}{\sqrt{n}} \min\left(\frac{c}{\sqrt{\sup p(x)}}, e^{-2}\sqrt{n} \right) \int \frac{p(x)}{h^d}\,dx.$$

Theorem 4 now follows from the fact that $\sup p(x) \to 0$ and that $\int (p(x)/h^d)\,dx = 1$.

REMARK. We leave the extension to general K as an exercise. When $f \in L_2(R)$, the integrated square error is relatively stable, that is, $\int (f_n - f)^2 / E(\int (f_n - f)^2) \to 1$ in probability, at least when h, f, and K satisfy some regularity conditions (Hall, 1982). In a milestone paper, Hall (1984) actually obtained the asymptotic law of $\int (f_n - f)^2$ when f has two bounded uniformly continuous derivatives on R^d, and K is a bounded density corresponding to a zero mean random vector with zero off-diagonal covariances, and unit variance components.

The techniques used in the proofs of Theorems 1 and 2 lead to useful results related to almost sure stability. This is illustrated below for the cubic histogram estimate.

THEOREM 5. *Let f_n be the cubic histogram estimate in R^d based on positive constants a_i, $1 \le i \le d$, and scaling factor h (notation of Section 3), where $\lim_{n\to\infty} h = 0$, $\lim_{n\to\infty} nh^d = \infty$. Let f be an arbitrary density on $[0,1]^d$, and let c be the constant of Lemma 8. Then, for all $\varepsilon \in (0,1)$ there*

exists $n_0 > 0$, *such that for* $n \geq n_0$,

$$P\left(\frac{J_n}{E(J_n)} \geq 1 + \frac{\sqrt{20}}{c\int\sqrt{f}(1-\varepsilon)} \right) \leq 3\exp\left(-\frac{4}{5}\left(\prod_{i=1}^{d} a_i h^d (1-\varepsilon) \right)^{-1} \right)$$

If also $\lim_{n\to\infty} h^d \log(n) = 0$, *then*

$$\limsup_{n\to\infty} \frac{J_n}{E(J_n)} \leq 1 + \frac{\sqrt{20}}{c\int\sqrt{f}} \quad \textit{almost surely}.$$

Proof. The rectangles defining the cubic histogram estimate, A_{nj}, $j = 1, \ldots$ have sides of lengths ha_i, $i = 1, \ldots, d$. The number of A_{nj}'s with $\mu(A_{nj}) > 0$ is denoted by N. Clearly,

$$N \leq \prod_{i=1}^{d} \left(2 + \frac{1}{ha_i} \right).$$

We define the constant b by $1 + \sqrt{20}/(c\int\sqrt{f}(1-\varepsilon))$. Now, by Lemmas 6 and 8, and using the notation of Sections 3 and 4, we have

$$\begin{aligned}
&P\left(J_n/E(J_n) \geq b\right) \\
&\quad \leq P\left(1 + 2\int |f_n - E(f_n)| \Big/ E\left(\int |f_n - E(f_n)| \right) \geq b \right) \\
&\quad = P\left(\sum_j |\mu_n(A_{nj}) - \mu(A_{nj})| \geq \frac{b-1}{2} \sum_j E\left(|\mu_n(A_{nj}) - \mu(A_{nj})| \right) \right) \\
&\quad = P\left(\sum_j |\mu_n(A_{nj}) - \mu(A_{nj})| \geq (b-1) \sum_j E\left(\left(\mu(A_{nj}) - \mu_n(A_{nj}) \right)_+ \right) \right) \\
&\quad \leq P\left(\sum_j |\mu_n(A_{nj}) - \mu(A_{nj})| \geq (b-1) \right. \\
&\qquad \left. \times \sum_j \min\left(e^{-2}\mu(A_{nj}); c\sqrt{\mu(A_{nj})/n} \right) \right).
\end{aligned}$$

But

$$\sum_j \min\left(e^{-2}\mu(A_{nj}); c\sqrt{\mu(A_{nj})/n}\right)$$

$$\geq \sum_j c\sqrt{\mu(A_{nj})/n} - \sum_{j:\,\mu(A_{nj})<c^2e^4/n} c\sqrt{\mu(A_{nj})/n}$$

$$\geq \sum_j c\int_{A_{nj}} \sqrt{f}\Big/\sqrt{n\lambda(A_{nj})} - \frac{Nc^2e^2}{n}$$

$$= \frac{c\int\sqrt{f}}{\sqrt{\prod a_i(nh^d)}} - O\left(\frac{1}{nh^d}\right).$$

Thus, since $nh^d \to \infty$, we see that for all n large enough,

$$P\left(J_n/E(J_n) \geq b\right) \leq P\left(\sum_j |\mu_n(A_{nj}) - \mu(A_{nj})| \geq \delta\right)$$

where

$$\delta = \left(\frac{20}{(\prod a_i)nh^d(1-\varepsilon)}\right)^{1/2}.$$

Next, note that $N/n \leq \delta^2/20$ for all n large enough (since $N/n \sim (nh^d\prod a_i)^{-1}$). Thus, by Lemma 1,

$$P\left(J_n/E(J_n) \geq b\right) \leq 3\exp\left(-n\delta^2/25\right),$$

which was to be shown. The last statement of Theorem 5 follows from this inequality and the Borel–Cantelli lemma.

REFERENCES

S. Abou-Jaoude (1976a). Sur une condition nécessaire et suffisante de L_1-convergence presque complète de l'estimateur de la partition fixe pour une densité, *Comptes Rendus de l'Académie des Sciences de Paris Série A* **283**, pp. 1107–1110.

S. Abou-Jaoude (1976b). Sur la convergence L_1 et L_∞ de l'estimateur de la partition aléatoire pour une densité, *Annales de l'Institut Henri Poincaré* **12**, pp. 299–317.

S. Abou-Jaoude (1976c). Conditions nécessaires et suffisantes de convergence L_1 en probabilité de l'histogramme pour une densité, *Annales de l'Institut Henri Poincaré* **12**, pp. 213–231.

S. Abou-Jaoude (1977). "La convergence L_1 et L_∞ de certains estimateurs d'une densité de probabilité," Thèse de Doctorat d'État, Université Paris VI, Paris.

T. Cacoullos (1966). Estimation of a multivariate density, *Annals of the Institute of Statistical Mathematics* **18**, pp. 178–189.

L. Devroye (1983). The equivalence of weak, strong and complete convergence in L_1 for kernel density estimates, *Annals of Statistics* **11**, pp. 896–904.

P. Hall (1982). Limit theorems for stochastic measures of the accuracy of density estimators, *Stochastic Processes and Applications* **13**, pp. 11–25.

P. Hall (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators, *Journal of Multivariate Analysis* **14**, xx–xx.

W. Hoeffding (1963). Probability inequalities for the sum of bounded random variables, *Journal of the American Statistical Association* **58**, pp. 13–30.

E. Parzen (1962). On the estimation of a probability density function and the mode, *Annals of Mathematical Statistics* **33**, pp. 1065–1076.

M. Rosenblatt (1956). Remarks on some nonparametric estimates of a density function, *Annals of Mathematical Statistics* **27**, pp. 832–837.

CHAPTER 4

Lower Bounds for Rates of Convergence

1. INTRODUCTION

In this chapter we will try to obtain general information about possible rates of convergence for $E(\int |f_n - f|)$, for all density estimates, in the form of lower bounds. There are two questions one can ask here:

(i) Obtain lower bounds for

$$\sup_{f \in \mathscr{F}} E\left(\int |f_n - f|\right),$$

where $\mathscr{F}$ is a suitably restricted class of densities. We will call these *uniform lower bounds*.

(ii) Obtain lower bounds for

$$\sup_{f \in \mathscr{F}} \limsup_{n \to \infty} a_n^{-1} E\left(\int |f_n - f|\right),$$

where a_n is a sequence of positive numbers. Thus, in (ii) we ask for the worst possible rate of convergence for a single density f in $\mathscr{F}$.

Occasionally, we will refer to the quantity $\inf_{f_n} \sup_{f \in \mathscr{F}} E(\int |f_n - f|)$ (which depends only upon n and $\mathscr{F}$) as the *minimax error*, and to lower and upper bounds for it as *minimax lower bounds* and *minimax upper bounds*.

The following classes of densities on R^1 will be considered:

G: all densities vanishing outside $[0, 1]$ and bounded by 2.

G_∞: all densities vanishing outside $[0, 1]$, bounded by $2 + \delta$ (some $\delta > 0$), and infinitely many times continuously differentiable on $[0, 1)$.

$H(g)$: all densities of the form $\sum_{i=1}^{\infty} p_i g(x + x_i)$, where g is an arbitrary fixed density with support contained in $[0, 1]$, $(p_1, p_2, \ldots)$ is a probability vector, and x_i is an increasing sequence of real numbers with $x_{i+1} - x_i > 1$.

U: all densities on $[0, \infty)$ that are monotone and have a peak at 0.

U_∞: all symmetric infinitely many times continuously differentiable unimodal densities with center at 0.

THEOREM 1. *Let f_n be any density estimate.*

(i) $$\inf_n \sup_{f \in \mathscr{F}} E\left(\int |f_n - f|\right) \geq c$$

where $c = 1$ for $\mathscr{F} = G$, G_∞ or $H(g)$, and $c = \frac{1}{8}$ for $\mathscr{F} = U$ or U_∞.

(ii) *Let $\{a_n\}$ be a sequence of positive numbers tending to 0. Then, for all $\mathscr{F}$ mentioned in* (i),

$$\sup_{f \in \mathscr{F}} \limsup_{n \to \infty} \frac{1}{a_n} E\left(\int |f_n - f|\right) = \infty.$$

Theorem 1 will be called a *slow convergence theorem*. To study rates of convergence for any density estimate, it is clearly not sufficient to put continuity conditions on $\mathscr{F}$ because Theorem 1 covers the classes U_∞ and $H(g)$. For example, if $g(x) = c\exp(-1/x(1-x))$, $0 \leq x \leq 1$, every f in $H(g)$ is infinitely many times continuously differentiable. Furthermore, because G is also included in Theorem 1, a tail condition alone or a boundedness condition alone does not suffice either. Thus, a combination of continuity and tail conditions seems necessary to obtain meaningful uniform and individual lower bounds. But even here, in view of G_∞, one must be careful: the nondifferentiability of f at even one point suffices to obtain a slow convergence result.

We note that in part (ii) of Theorem 1, *one* f is chosen in $\mathscr{F}$: it does usually depend upon the sequences f_n and a_n, but once chosen, remains the same for all n. In part (ii) we have proved that any rate of convergence is attainable within the subclasses $\mathscr{F}$ considered here. We should note here that in part (ii) the lim sup can be replaced by a lim inf (Birgé, 1983b).

Theorem 1 (i) is not totally satisfying in that the particular densities *within* the subclasses $\mathscr{F}$ that give us the large values for $E(\int |f_n - f|)$ are possibly those f that correspond to large values of a criterion that measures how long-tailed or un-smooth f is. One such criterion that will reappear in a

natural way in Chapter 5 is

$$D_s(f) = \left(\int |f^{(s)}|\right)^{1/(2s+1)} \left(\int \sqrt{f}\right)^{2s/(2s+1)},$$

where $f^{(s)}$ is the sth derivative of f. Note here that $\int\sqrt{f}$ measures the heaviness of the tail of f. The integer s can take any positive value. We will now see that Theorem 1 (i) is largely due to the presence within each $\mathcal{F}$ of densities with large values of $D_s(f)$. In doing so, we will only consider uniform lower bounds.

THEOREM 2. *Let f_n be any density estimate. Let g be any density on* $[0,1]$ *with continuous* sth *derivative* $g^{(s)}$. *Then*

$$\liminf_{n\to\infty} n^{s/(2s+1)} \sup_{f\in H(g)} \frac{E\left(\int |f_n - f|\right)}{D_s(f)} \geq \frac{(s/e(2s+1))^{s/(2s+1)}}{D_s(g)}, \quad \text{all } s \geq 1.$$

We note here that $\inf_g D_s(g) = C(s) > 0$ for $s = 1, 2$ (see Chapter 5), so that the lower bound is

$$\frac{(s/e(2s+1))^{s/(2s+1)}}{C(s)}, \qquad s = 1, 2,$$

provided we replace the supremum over all f in $H(g)$ by that over all g on $[0,1]$ and all f in $H(g)$.

It is clear that the supremum in Theorem 2 is not approached by densities that have $D_s(f) = \infty$, as may have been the case in Theorem 1. Thus, Theorem 2 tells us about the worst f that are in a sense reasonably well-behaved (because $D_s(f) < \infty$). Theorem 2 also highlights the importance of the normalizing factor.

To illustrate Theorem 2 in the case $s = 2$, we will jump ahead a bit, and announce a result of Chapter 5 for kernel estimates f_n: for all f, we must have

$$\liminf_{n\to\infty} n^{2/5} \frac{E\left(\int |f_n - f|\right)}{D_2(f)} \geq c > 0,$$

where c is a universal constant, and $D_2(f)$ is defined as above when f is

twice continuously differentiable (its definition is different in the other cases). This has information not available from Theorem 2 because it is an individual (not a uniform) result. Yet, it implies that, for $s \geq 3$, the lower bound of Theorem 2 is not achievable with the kernel estimate. For $s \geq 3$, we need either other estimates or a drastically modified kernel estimate. In Sections 5.9 and 7.6 we will see that it suffices to allow negative-valued K.

A careful analysis of the proof of Theorem 2 reveals that for the densities f_n^* in $H(g)$ for which $E(\int |f_n - f|)/D_s(f)$ is large, $D_s(f_n^*) \to \infty$. We could take the bounding technique a little further now by restricting ourselves to a class of well-behaved densities such as

$F_{s,r}$: all densities on $[0, 1]$ with $(s-1)$ absolutely continuous derivatives, sth derivative $f^{(s)}$, and $D_s(f) \leq r$,

or

$F_{s,\infty}$: all densities on $[0, 1]$ with $(s-1)$ absolutely continuous derivatives, sth derivative $f^{(s)}$, and $D_s(f) < \infty$.

These classes are not nested with respect to any of the classes considered until now. We have another theorem:

THEOREM 3 (Bretagnolle and Huber, 1979). *Let r^* be a number at least equal to*

$$\left(\tfrac{1}{4}9^s(s+1)!\right)^{1/(2s+1)}.$$

For any density estimate f_n, we have

$$\liminf_{n\to\infty} n^{s/(2s+1)} \sup_{f\in F_{s,r}} E\left(\int |f_n - f|\right) \geq (2e)^{-4}\left(\frac{r}{r^*} - 1\right), \quad \text{all } r > r^*,$$

and

$$\liminf_{n\to\infty} n^{s/(2s+1)} \sup_{f\in F_{s,\infty}} E\left(\int |f_n - f|\right) = \infty.$$

Theorem 3 is stronger than Theorem 2 in the sense that the suprema are taken over classes of densities $F_{s,r}$ having uniformly bounded values of $D_s(f)$. As a consequence, the argument is more subtle and sophisticated. We observe again that the rate $n^{-s/(2s+1)}$ approaches $n^{-1/2}$ as $s \to \infty$. While this rate is not achievable with the kernel estimate for $s \geq 3$ for any f, it is achievable at least on an individual basis with some other density estimates for densities f in $F_{s,\infty}$ having compact support: see, for example, Bartlett's estimate described in Section 7.6 or Section 5.9.

All of the classes treated thus far are still quite large. Further drastic reductions in the sizes of the classes will of course result in smaller lower bounds. If we reduce the classes to such an extent that there is only one parameter θ left in the family, our lower bounds should be valid for all parametric density estimates for the given class. The lower bounds thus obtained will usually not be attainable by the general density estimates considered here. They should however be tight for some specific parametric density estimates. For example, consider the following simple family of densities:

$\Pi(g)$: all densities of the form $f(x) = pg(x) + (1-p)g(x+2)$, where g is an arbitrary density with support contained in $[0,1]$, and $p \in [0,1]$ is a mixing parameter, unknown to the user.

Note that all f in $\Pi(g)$ have compact support, and have infinitely many continuous derivatives when g has. We will prove the following theorem:

THEOREM 4. *Let f_n be any density estimate and let g be an arbitrary density with support contained in* $[0,1]$. *Then,*

(i) *For all $n \geq 4$,*

$$\sup_{f \in \Pi(g)} \sqrt{n}\, E\left(\int |f_n - f|\right) \geq 0.030153\cdots$$

$$\left(\textit{in fact}, \sup_{f \in \Pi(g)} E\left(\int |f_n - f|\right) \geq (0.0849856\cdots + o(1))/\sqrt{n}\right).$$

(ii) $$\sup_{f \in \Pi(g)} \limsup_{n \to \infty} \sqrt{n}\, E\left(\int |f_n - f|\right) \geq 0.0424928\cdots.$$

We can thus conclude that unless one chooses a truly trivial class $\mathcal{F}$, the best possible rate of convergence in L_1 is $1/\sqrt{n}$. For example, if $\mathcal{F}$ has only a finite number of members, $g_1, \ldots, g_N$, and we define $f_n = g_i$, where i is determined after having looked at the data $X_1, \ldots, X_n$, we have $E(\int |f_n - f|) \leq 2P(g_i \neq f)$, and this tends to 0 exponentially fast in n (see Section 11.9 on detection) when g_i is chosen by the maximum likelihood principle. Thus, *for all f in this finite $\mathcal{F}$, we have an exponentially* decreasing *upper* bound!

A common complaint of users is that most statistical theories are asymptotic in nature. What can one do for small samples? Is there such a thing as "small sample superperformance"? Well, one way of obtaining good estimates for small n consists of suitably restricting the densities f one

is willing to consider, and of answering the following minimax question: for which f_n is $\sup_{f \in \mathcal{F}} E(\int |f_n - f|)$ minimal? (This minimal value, $m(n, \mathcal{F})$, is a function of n and $\mathcal{F}$ only.) Classes one might consider here are all monotone densities on $[0, \infty)$; all symmetric unimodal densities; all Lipschitz (C) densities; all log concave densities with mode at 0; all densities with increasing hazard rate on $[0, \infty)$; and so on. In particular, if only X_1 is given (the *one-observation-problem*), what is f_1 for some of these classes? It is of interest of course to have good bounds for $m(n, \mathcal{F})$ for all n. Asymptotically good lower bounds can be obtained usually by the methods developed in this section and the next section, and are of less practical interest in this context.

2. ASSOUAD'S LEMMA

The principles used in Section 1 can be classified into three groups: first, one cannot estimate a density f in a given interval when no X_i falls in this interval (Theorems 1 and 2); secondly, lower bounds can be obtained by information-theoretic considerations (Theorem 3); thirdly, one can use inequalities such as the Cramér–Rao inequality, and properties of sufficient statistics (Theorem 4). In this section we would like to draw the attention to a powerful and simple technique developed by Assouad (1983) and Birgé (1980, 1983a, b), which will allow us to rederive some of the results of Section 1 and to obtain some new lower bounds for important classes of densities. Birgé used the notions of ϵ-entropy and ϵ-capacity introduced by Kolmogorov and Tikhomirov (1961) which allows him not only to obtain uniform lower bounds but also upper bounds for the minimax error

$$\inf_{f_n} \sup_{f \in \mathcal{F}} E\left(\int |f_n - f|\right).$$

He successfully answered the question of obtaining for some $\mathcal{F}$ lower and upper bounds that have the same dependency upon n (but different coefficients). Because we will obtain important upper bounds in Section 5, we will not concern ourselves with the second half of Birgé's work.

The key is a powerful lemma due to Assouad (1983) (Theorem 5 below), which we shall give in a form slightly adapted to our subject.

THEOREM 5 (Assouad, 1983) (General Form). *Let $r \geq 1$ be an integer, and let $b = (b_1, \ldots, b_r) \in \{-1, 1\}^r$ be a parameter of a family of densities*

$f(b,.)$ with 2^r members. Let b_{i+} and b_{i-} be the parameters defined by

$$b_{i+} = (b_1, b_2, \ldots, b_{i-1}, +1, b_{i+1}, \ldots, b_r),$$

$$b_{i-} = (b_1, b_2, \ldots, b_{i-1}, -1, b_{i+1}, \ldots, b_r).$$

If there exists a partition $A_0, A_1, \ldots, A_r$ of R^d such that for all b and all $1 \le i \le r$, the following inequalities are valid:

$$\int_{A_i} |f(b_{i+},.) - f(b_{i-},.)| \ge \alpha > 0$$

and

$$\int \sqrt{f(b_{i+},.) f(b_{i-},.)} \ge \beta > 0,$$

then, for any density estimate f_n,

$$\sup_b E\left(\int |f_n - f|\right) \ge \begin{cases} (r\alpha/2)\left(1 - \sqrt{2 - 2\beta^n}\right); \\ (r\alpha/4)\beta^{2n}. \end{cases}$$

In terms of $\gamma = 1 - \beta$, the lower bounds can be replaced by $(r\alpha/2)(1 - \sqrt{2n\gamma})$ and $(r\alpha/4)(1 - \gamma)^{2n}$, respectively.

(Particular Form) (Birgé, 1983b). *Let $r \ge 1$ be an integer, let $A = [0, l)$, $l \le 1/r$ be an interval on which we define a measurable function g having the properties*

$$|g| \le 1, \int_A g = 0,$$

let $g = 0$ outside A, and let $y_1, \ldots, y_r$ be real numbers such that the sets $A + y_i$ are nonoverlapping. Let f_0 be a density on R taking the value 1 on $\cup_i A + y_i$. Let F be the class of 2^r densities parametrized by

$$b = (b_1, \ldots, b_r) \in \{-1, 1\}^r$$

with members $f(b,.)$ defined as follows:

$$f(b,x) = \begin{cases} f_0(x), & x \notin \cup_i A + y_i; \\ f_0(x) + b_i g(x - y_i), & x \in A + y_i. \end{cases}$$

Then, for any *density estimate* f_n,

$$\sup_{f \in F} E\left(\int |f_n - f|\right) \geq r \int_A |g| \left(1 - \sqrt{2n \int_A g^2}\right)$$

$$\geq \tfrac{1}{2} r \int_A |g| \quad \left(\text{when } n \int_A g^2 \leq \tfrac{1}{8}\right).$$

If a general family $\mathscr{F}$ is given to us, we should first find r, l, g, f_0, and $y_1, \ldots, y_r$ so that the family F of Theorem 5 is entirely contained in $\mathscr{F}$. Then, any lower bound obtained over F is necessarily a lower bound over $\mathscr{F}$, and we are done. This plan of attack will now be illustrated for a few important classes.

We define the *Lipschitz class* $W(s, \alpha, C)$ as the class of all densities f vanishing outside $[0, 1]$, possessing $(s - 1)$ absolutely continuous derivatives and satisfying the condition

$$|f^{(s)}(x) - f^{(s)}(y)| \leq C|x - y|^{\alpha}, \qquad x, y \in R,$$

where $s \geq 0$ is an integer, C is a positive number, and $\alpha \in (0, 1]$. When we say that a function g is *Lipschitz* (C), we mean

$$|g(x) - g(y)| \leq C|x - y|, \qquad x, y \in R.$$

The analysis in Section 1 centered around the functionals $D_s(f)$. Here we start from a nice small class, the Lipschitz class $W(s, \alpha, C)$, without worrying at first about $D_s(f)$. We will show in Chapter 5 that uniformly over $W(s, 1, C)$, a suitably generalized functional related to $D_{s+1}(f)$ is uniformly bounded.

The most important Lipschitz classes are $W(1, 1, C)$ and $W(0, 1, C)$. The latter class is the class of all *Lipschitz* (C) densities that vanish outside $[0, 1]$. Obviously, $W(0, 1, C)$ is empty for $C < 4$, and has only one member (the isosceles triangular density on $[0, 1]$) when $C = 4$. Similarly, for all $W(s, \alpha, C)$, we observe that the class is nonempty for all C larger than a constant C_0 and empty for all $C < C_0$.

THEOREM 6. *Let* f_n *be an arbitrary density estimate. For all* $s \geq 0$ *(integer),* $\alpha \in (0, 1]$, *there exist positive constants* c_1, c_2, c_3, γ_1, *and* γ_2

depending upon s and α only *such that for all* $C \geq c_1$

$$\sup_{f \in W(s,\alpha,C)} E\left(\int |f_n - f|\right)$$

$$\geq \begin{cases} \left(c_3 + o(1)\right) C^{1/(2(s+\alpha)+1)} n^{-(s+\alpha)/(2(s+\alpha)+1)}; \\ \gamma_1 \dfrac{C}{4}\left(\left(16\gamma_2 n C^2\right)^{1/(2(s+\alpha)+1)} + 4\right)^{-(s+\alpha)}, \qquad n \geq Cc_2. \end{cases}$$

If $W^*(s,\alpha) = \cup_{C>0} W(s,\alpha,C)$, *then*

$$\liminf_{n\to\infty} \sup_{f \in W^*(s,\alpha)} E\left(\int |f_n - f|\right) n^{(s+\alpha)/(1+2(s+\alpha))} = \infty.$$

The constants c_2 and c_3 can be computed as follows:

$$\gamma_0 = \left[(s+\alpha)^{s+1} 2^{1-\alpha} \exp\left(\frac{s+\alpha}{2} \exp\left(\frac{s+1}{s+\alpha}\right)\right)\right]^{-1},$$

$$\gamma_1 = \gamma_0 \Gamma^2(s+\alpha+1)\Gamma^{-1}(2s+2\alpha+2),$$

$$\gamma_2 = \gamma_0^2 \Gamma^2(2s+2\alpha+1)\Gamma^{-1}(4s+4\alpha+2),$$

$$\gamma_3 = \gamma_0 / 4^{s+\alpha},$$

$$c_2 = \gamma_3^{(2(s+\alpha)+1)/(s+\alpha)} (16\gamma_2)^{-1},$$

$$c_3 = \tfrac{1}{4}\gamma_1 (16\gamma_2)^{-(s+\alpha)/(1+2(s+\alpha))}.$$

Theorem 6 is nice in that it provides us with a continuous range of polynomial rates of convergence. Unfortunately, the constants c_i in the theorem are suboptimal, and it is worthwhile to spend some extra effort on $W(1,1,C)$ and $W(0,1,C)$ in the hope of obtaining useful lower bounds. This is done in Theorem 7 below.

THEOREM 7. *Let* f_n *be any density estimate. Then, for all* $C \geq 72$,

$$\sup_{f \in W(0,1,C)} E\left(\int |f_n - f|\right) \geq \frac{21}{160}\left(\frac{12C}{25n}\right)^{1/3}, \qquad n \geq \operatorname{Max}\left(10, \frac{3C}{50}\right).$$

For all $C \geq 288$,

$$\sup_{f \in W(1,1,C)} E\left(\int |f_n - f|\right) \geq \begin{cases} \left(\frac{1}{32}\left(\frac{30}{23}\right)^{2/5} C^{1/5} + o(1)\right) n^{-2/5}, \\ \frac{C}{32}\left(\left(\frac{23nC^2}{30}\right)^{1/5} + 8\right)^{-2}, \qquad n \geq \frac{15C}{368}. \end{cases}$$

We will see in Chapter 5 that these rates are achievable by the kernel estimate, and, for $W(0,1,C)$ only, by the histogram estimate. If the lower bounds for C seem unrealistic, the user can without a lot of effort obtain smaller bounds at the expense of increased values of the coefficients of $n^{-1/3}$ and $n^{-2/5}$ in Theorem 7.

Let us show now that from Theorem 5, we can also obtain slow convergence theorems and fast convergence theorems in the spirit of Theorems 1 and 4.

THEOREM 8. *Let f_n be any density estimate. Let $r \geq 1$ be a fixed integer, and let g be a fixed measurable function on $[0,1/r)$ with $|g| \leq 1$, $\int_0^{1/r} g = 0$. Let $Q_r(g)$ be the class of all densities of the following form: there exists an $\varepsilon \in [0,1]$ and numbers $b_i \in \{-1,1\}$, $1 \leq i \leq r$, such that*

$$f(x) = 1 + \varepsilon b_i g\left(x - \frac{i}{r}\right), \qquad \frac{i}{r} \leq x < \frac{i+1}{r}, \qquad i = 0,1,\ldots,r-1.$$

Then,

$$\sup_{f \in Q_r(g)} E\left(\int |f_n - f|\right) \geq \frac{r\int |g|}{\sqrt{32n\int g^2}}, \qquad n \geq \frac{1}{\left(8\int g^2\right)}.$$

In particular, if $g = 1$ on $[0,1/(2r))$ and $g = -1$ on $[1/(2r),1/r)$, then

$$\sup_{f \in Q_r(g)} E\left(\int |f_n - f|\right) \geq \sqrt{r/32n}, \qquad n \geq r/8,$$

and

$$\sup_{f \in U_r Q_r(g)} E\left(\int |f_n - f|\right) \geq \frac{1}{2} \quad \text{for all } n.$$

Theorem 8 contains quite a bit of information despite its simplicity. The

uniform lower bound over $Q_1(g)$ is of the order of $1/\sqrt{n}$. This lower bound applies to the *parametric* estimation problem because it remains valid even if g were given to us, but only ε and the b_i's were unknown. Thus, we are once again in a situation comparable to that of Theorem 4. Not surprisingly, the lower bound grows with r, and by considering $\cup_r Q_r(g)$ we are in fact back in the domain of Theorem 1. This result comes as no surprise because $\cup_r Q_r(g)$ is a densely populated subclass of G, the class of all densities on $[0,1]$ that are bounded by 2.

All results of this section were uniform results. Individual lower bounds as in Theorem 1 (ii) can also be obtained from Theorem 5 by involved constructions (see, e.g., Birgé, 1983b).

We close this section with a uniform lower bound for the class $M_B = \{\text{all nonincreasing densities on } [0,1] \text{ with } f(0) \le B\}$. It is clear that this class is empty unless $B \ge 1$.

THEOREM 9. *For any density estimate and any $B \ge 2$,*

$$\sup_{f \in M_B} E\left(\int |f_n - f|\right) \ge \frac{1}{16\left(3 + 2(n/4)^{1/3}\right)} \sim \frac{1}{32}\left(\frac{4}{n}\right)^{1/3}.$$

Clearly, the same lower bound is valid for all symmetric unimodal densities on $[-1,1]$ satisfying $f(0) \le B/2$, and for all unimodal densities on $[-1,1]$ with mode at $m \in [-1,1]$ and $f(m) \le B/2$. The lower bound developed here is far from best possible: for one thing, it is not an increasing function of B. The proof of Theorem 9 however is amazingly simple, and at least the power of n ($n^{-1/3}$) is correct, because we will see in Chapter 5 that both the kernel and the histogram estimate have uniform upper bounds for the expected L_1 error which increase as $n^{-1/3}$. The lesson we learn from this is that under monotonicity and compactness conditions alone, no grand performances should be expected from any estimate, and that it is probably not very rewarding to construct special estimates for M_B, possibly not consistent outside M_B, since little can be gained over the kernel estimate.

IMPORTANT REMARK. To obtain uniform lower and upper bounds for $E(\int |f_n - f|)$ over nonhomogeneous or too large classes $\mathcal{F}$ is a risky and often useless exercise. It is much like trying to determine the maximal number of worms in a single apple in a bunch of freshly picked apples after having thrown in a couple of old apples. With high probability, the old apples will determine the outcome, and we end up with little or no information about the freshly picked apples. One instance of this phenomenon occurs for Lipschitz classes $W(s, \alpha, C)$, where the upper and lower

bounds increase with C (for fixed s and α). From Theorems 2 and 3, and from upper bounds to be derived in Chapter 5, we will see that the quantity $D_s(f)$ measures the "difficulty" posed by f very well. But because C and $D_s(f)$ are only vaguely correlated, we obtain very little information about the vast majority of the densities in $W(s, \alpha, C)$ from Theorem 6, say. In that respect, the classes $F_{s,r}$, which from now on we shall call *Bretagnolle–Huber* classes, seem more natural and realistic. In addition, $W(s, \alpha, C)$ is not closed under rescaling, while $F_{s,r}$ is.

3. SOME HISTORICAL REMARKS

The collection of lower bounds of Section 1 has some L_p analogues. For obvious reasons, this is not the forum for such lower bounds. It is of interest however to recall some historical milestones because they will help us to better understand the difference between the L_1 norm and the L_p norms for $p \neq 1$.

In Section 1, we considered slow convergence results (Theorem 1), medium rate lower bounds (Theorems 2 and 3), and small lower bounds (Theorem 4). Theorem 1 leads, for example, to an important observation about L_2 convergence results for the kernel estimate,

$$f_n(x) = (nh_n)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right), \qquad x \in R,$$

for bounded symmetric densities K. Rosenblatt (1971) has shown that

$$E\left(\int (f_n - f)^2\right) \sim \frac{\alpha}{nh_n} + \frac{\beta}{4} h_n^4$$

when $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$, and f is a bounded density with two continuous derivatives and $\int f''^2 < \infty$. The constants are

$$\alpha = \int K^2, \qquad \beta = \left(\int x^2 K(x)\, dx\right)^2 \int f''^2.$$

Taking $h_n = (\alpha/\beta n)^{1/5}$ yields the optimal L_2 rate

$$E\left(\int (f_n - f)^2\right) \sim \frac{5}{4} \frac{\alpha^{4/5} \beta^{1/5}}{n^{4/5}}$$

(see also Nadaraya, 1974). Yet, at the same time, for some f in the given

class,

$$E\left(\int |f_n - f|\right) \geq \frac{1}{\log\log\log\log n}$$

infinitely often: in Theorem 1, take $H(g)$ with $g(x) = \text{constant} \cdot \exp(-1/x(1-x))$ on $[0,1]$, and note that for all f in $H(g)$, f is bounded and $\int f''^2 = \sum_{i=1}^{\infty} p_i^2 \int g''^2 \leq \int g''^2 < \infty$. In other words, Rosenblatt's classical result and most other L_2 rate of convergence results give us little information about how close f_n is to f, and should be used with extreme care when it comes to choosing h_n. This discrepancy between good L_2 rates and poor L_1 rates is due to the fact that in L_2, tails (and regions with low-f values) are less important. We should note though that if some tail conditions are added to the class of densities considered by Rosenblatt, the optimal L_1 rate for the kernel estimate ($n^{-2/5}$) is achieved (see Chapter 5).

We note without proof that Theorem 1 has an L_p analogue.

THEOREM 10 (Devroye, 1983). *Let f_n be any density estimate, let $p \geq 1$ be a fixed real number, and let $f \in L_p$. Then,*

(i)
$$\inf_n \sup_{f \in \mathcal{F}} \frac{E\left(\int |f_n - f|^p\right)}{\int f^p} \geq \frac{1}{2^{p-1}}$$

where $\mathcal{F} = H(g)$ for any g, or $\mathcal{F} = G$.

(ii) *Let $\{a_n\}$ be a sequence of positive numbers tending to 0. Then*

$$\sup_{f \in G} \limsup_{n \to \infty} a_n^{-1} \frac{E\left(\int |f_n - f|^p\right)}{\int f^p} = \infty.$$

There are also several analogues for the medium rate lower bounds, cited here without proof, or with a short sketch of proof only. We remind the reader that it is easy to get lost in the vast sea of results available in the literature: there are as many results as there are normalization factors, norms, and classes of densities $\mathcal{F}$. In what follows, we will give a generalization of the factor $D_s(f)$ toward L_p: we will define

$$D_{sp}(f) = \left(\int |f^{(s)}|^p\right)^{1/(2s+1)} \left(\int f^{p/2}\right)^{2s/(2s+1)}$$

The uniform lower bounds of Theorem 1 and the uniform lower bound of

Theorem 3 are complemented by the following result by Bretagnolle and Huber (1979), valid for $d = 1$, $s \geq 1$, and $p > 1$:

$$\liminf_{n \to \infty} n^{sp/(2s+1)} \sup_{f \in \mathscr{F}} E\left(\int |f_n - f|^p\right) \geq rC_{sp} > 0,$$

where C_{sp} is a constant depending upon s and p only, $r > 0$ is a real number, and $\mathscr{F}$ is the class of all densities with $(s-1)$ absolutely continuous derivatives, $f^{(s)} \in L_p$, $f \in L_p$, and $D_{sp}(f) \leq r$. We note that for $p \geq 2$, the condition $D_{sp}(f) \leq r$ does not impose restrictions on the tail of f, and thus this result does not contradict Rosenblatt's L_2 rate of convergence result for kernel estimates.

Let us stress once more the importance of normalizations. For example, for $p \geq 2$, we can find g such that $H(g) \subseteq \mathscr{F}$: this follows from the fact that for $f \in H(g)$,

$$D_{sp}(f)^{2s+1} = \left(\sum p_i^p\right)\left(\sum p_i^{p/2}\right)^{2s} D_{sp}(g)^{2s+1} \leq D_{sp}(g)^{2s+1}.$$

Thus, by Theorem 9 $\sup_{f \in \mathscr{F}} E(\int |f_n - f|^p)/\int f^p \geq 1/2^{p-1}$, all n, all $p \geq 2$.

Let us now gradually reduce the size of the classes $\mathscr{F}$. First consider $H(g)$ again. A check of Section 5 below reveals that

$$\sup_{f \in H(g)} E\left(\int |f_n - f|^p\right) \geq \int g^p 2^{-(p-1)} \sup_{\substack{\text{all probability} \\ \text{vectors } p_1, p_2, \ldots}} \sum_{i=1}^{\infty} p_i^p (1 - p_i)^n$$

$$\geq \int g^p \left(\frac{p-1}{2(1+n)}\right)^{p-1} \left(1 - \frac{p-1}{1+n}\right)^n$$

$$\geq \int g^p \left(\frac{(p-1)}{2(n+1)} \exp\left(-\frac{n}{2+n-p}\right)\right)^{p-1}$$

$$\sim \int g^p \frac{((p-1)/2)^{p-1}}{n^{p-1}} \quad \text{as } n \to \infty.$$

The lower bound is 1 for $p = 1$ and all n. For $p = 2$, it reduces to a constant times $1/n$, regardless of the choice of g. In fact, since $\int g^p \geq 1$ for all $p \geq 1$, we have, for all g,

$$\sup_{f \in H(g)} E\left(\int (f_n - f)^2\right) \geq \frac{1}{2(n+1)} \exp\left(-\frac{n}{1+n}\right), \quad \text{all } n.$$

Yet another normalization yields the following result.

THEOREM 11. *Let g be an infinitely many times continuously differentiable density with support on $[0,1]$, and let $d = 1$. Then*

$$\liminf_{n\to\infty} n^{ps/(2s+1)} \sup_{f\in H(g)} \frac{E\left(\int |f_n - f|^p\right)}{D_{sp}(f)} \geq \frac{1}{2^{p-1}} \frac{\left(\frac{ps}{2s+1}\cdot\frac{1}{e}\right)^{ps/(2s+1)}}{D_{sp}(g)},$$

for all $p, s \geq 1$, and all density estimates f_n.

We have seen that $1/n$ is the best L_2 rate of convergence attainable uniformly over any class $H(g)$. Boyd and Steele (1978) have shown something quite a bit stronger for the small class

BS: all normal $(0, \sigma^2)$ densities.

THEOREM 12 (Boyd and Steele, 1978). *For any density estimate f_n, there exists an $f \in BS$ such that*

$$\limsup_{n\to\infty} nE\left(\int (f_n - f)^2\right) \geq c(f) > 0,$$

where $c(f)$ is a constant depending only upon f.

The result of Boyd and Steele cannot be improved for normal density estimation: for example, if we estimate f by f_n, a normal $(\hat{\mu}, \hat{\sigma}^2)$ density where $\hat{\mu}$ and $\hat{\sigma}^2$ are the standard sample-based estimates of μ and σ^2, then

$$\lim_{n\to\infty} P\left(\int (f_n - f)^2 < \frac{x}{16\sqrt{\pi}\, n\sigma}\right) = F(x),$$

where F is the distribution function of $4V + 3U$, and V, U are independent chi-square random variables with one degree of freedom (see Maniya, 1969, who also has a similar result for $d > 1$). Thus, the rate predicted by Boyd and Steele can be achieved. Theorem 12 should be considered as the L_2 counterpart of Theorem 4. It is conjectured that

$$\sup_{f\in BS} \limsup_{n\to\infty} \sqrt{n}\, E\left(\int |f_n - f|\right) \geq c > 0$$

for some universal constant c.

Kiefer (1982) surveyed the literature on lower bounds for rates of convergence in density estimation: until now, most emphasis has been on pointwise rates of convergence (see Farrell 1967, 1972; Wahba, 1975; Stone, 1980; Ibragimov and Khasminskii, 1981). Global or L_p rates of convergence

have been considered by Samarov (1977) and Bretagnolle and Huber (1979). For L_∞ lower bounds, see, for example, Ibragimov and Khasminskii (1981, Theorem 4.2) and Stone (1983). For the kernel estimate, the rate of pointwise convergence was studied in some detail by Wahba (1975), Rosenblatt (1971), and many others. Its L_2 rate of convergence was obtained by Rosenblatt (1971), Nadaraya (1974), Deheuvels (1977a), and Bretagnolle and Huber (1979), among others. For $d > 1$, we refer to Deheuvels (1977b).

4. PROOF OF THEOREM 1

Classes G and $H(g)$. We will start with two families of densities. Family 1 is parametrized by a real number $b \in [0,1]$ and a probability vector $(p_1, p_2, \ldots)$. We will only make the dependence upon b explicit and write

$$f(b,x) = \sum_{i=1}^{\infty} p_i g(x - 2i - b_i),$$

where g is a density with support contained in $[0,1]$, and b_i is the ith bit in the binary expansion of $b = 0.b_1b_2b_3\cdots$. For each b, $f(b,x)$ is a density in x.

Family 2 is parametrized by the same b and the same probability vector $(p_1, p_2, \ldots)$. First we partition $[0,1]$ into sets A_i, A_i' such that $\int_{A_i} dx = \int_{A_i'} dx = p_i/2$, and then we define

$$f(b,x) = 2\sum_{i=1}^{\infty} I_{b_iA_i + (1-b_i)A_i'}(x).$$

We can always take A_i and A_i' as adjacent intervals such that $A_i' = A_i + p_i/2$. It is clear that family 2 is a subclass of G, and that family 1 is a subclass of $H(g)$.

The proof is based upon the following embedding device: let B be a uniform $[0,1]$ random variable, and let $X_1^*, \ldots, X_n^*$ be independent random variables, independent of B, with common density $f(0,.)$. Then define the sets $C_i = [2i, 2i+2)$ for family 1, and $C_i = A_i \cup A_i'$ for family 2. Clearly, $P(X_1^* \in C_i) = p_i$ in both cases. We will also use the notation B_i for the ith bit in the binary expansion of B. The random variables $X_1, \ldots, X_n$ are now defined by the relations

$$X_j = X_j^* + B_i \quad \text{when } X_j^* \in C_i \quad \text{(family 1)};$$

$$X_j = X_j^* + \frac{B_i p_i}{2} \quad \text{when } X_j^* \in C_i \quad \text{(family 2)}.$$

Furthermore, we will use the random variables $N_i = \sum_{j=1}^{n} I_{[X_j \in C_i]} = \sum_{j=1}^{n} I_{[X^* \in C_i]}$, so that, by construction, $N = (N_1, N_2, \ldots)$ is *independent of* B. Finally, we will let B' and B'' be random variables equal to B except in their ith digits, where we force $B_i' = 0$, $B_i'' = 1$.

There are two fundamental tricks: the first trick is based upon the fact that a supremum is larger than an expected value:

$$\sup_b E\left(\int |f_n - f(b,.)|\right) \geq E\left(\int |f_n - f(B,.)|\right)$$

$$= E\left(\sum_{i=1}^{\infty} \int_{C_i} |f_n - f(B,.)|\right). \tag{1}$$

The second trick eliminates f_n from (1) and uses the fact that on $[N_i = 0]$, B_i and $X_1, \ldots, X_n$ are conditionally independent. Thus,

$$E\left(I_{[N_i=0]} \int_{C_i} |f_n - f(B,.)| \,\|\, X_1, \ldots, X_n\right)$$

$$= E\left(I_{[N_i=0]} I_{[B_i=0]} \int_{C_i} |f_n - f(B',.)|\right.$$

$$\left. + I_{[N_i=0]} I_{[B_i=1]} \int_{C_i} |f_n - f(B'',.)| \,\|\, X_1, \ldots, X_n\right)$$

$$= \frac{1}{2} I_{[N_i=0]} E\left(\int_{C_i} |f_n - f(B',.)| + \int_{C_i} |f_n - f(B'',.)| \,\|\, X_1, \ldots, X_n\right)$$

$$\geq I_{[N_i=0]} \frac{1}{2} E\left(\int_{C_i} |f(B',.) - f(B'',.)|\right)$$

$$= p_i I_{[N_i=0]} \tag{2}$$

for both families. A combination of (1) and (2) now gives

$$\sup_b E\left(\int |f_n - f(b,.)|\right) \geq \sum_{i=1}^{\infty} p_i P(N_i = 0) = \sum_{i=1}^{\infty} p_i (1 - p_i)^n. \tag{3}$$

Inequality (3) is strong enough to prove (i) for families G and $H(g)$, because we still have the freedom to choose $(p_1, p_2, \ldots)$. Consider, for example, $p_i = 1/M$, $1 \leq i \leq M$, and $p_i = 0$, $i > M$. Then (3) is equal to

$(1 - 1/M)^n$, and the supremum over all M is 1. We will postpone the classes G_∞, U, and U_∞ for the time being.

For part (ii) we will need a Lemma.

LEMMA 1. *When* $0 < a_n \le \frac{1}{8}$ *for all* n, $\lim_{n\to\infty} a_n = 0$, *then there exists a probability vector* $(p_1, p_2, \ldots)$ *such that*

$$\sum_{i=1}^{\infty} p_i(1 - p_i)^n \ge a_n, \quad \text{all } n.$$

Proof. Construct first the sequence

$$a_n' = \max_{m \ge n} a_m + \frac{1}{4(n+1)}.$$

This sequence satisfies: $a_n' \ge a_n$, for all n, $\frac{1}{4} \ge a_1' \ge a_2' \ge \cdots \ge a_n' \downarrow 0$. Thus, we can find integers $1 = k_1 < k_2 < \cdots$ and positive numbers p_i such that $p_1 = 1 - 2a_1'$, and for $n \ge 2$, $k_{n-1} < i \le k_n$:

$$p_i \le (2n)^{-1}, \qquad \sum_{i=k_{n-1}+1}^{k_n} p_i = 2(a_{n-1}' - a_n').$$

Note in particular that

$$\sum_{i=1}^{\infty} p_i = p_1 + \sum_{n=2}^{\infty} 2(a_{n-1}' - a_n') = 1 - 2a_1' + 2a_1' = 1.$$

Also, for $n \ge 2$,

$$\begin{aligned}\sum_{i=1}^{\infty} p_i(1 - p_i)^n &\ge \left(1 - \frac{1}{2n}\right)^n \sum_{p_i \le 1/2n} p_i \\ &\ge \frac{1}{2} \sum_{p_i \le 1/2n} p_i \ge \frac{1}{2} \sum_{i=k_{n-1}+1}^{\infty} p_i \\ &= \frac{1}{2} \sum_{i=n}^{\infty} 2(a_{i-1}' - a_i') = a_{n-1}' \ge a_n' \ge a_n.\end{aligned}$$

For $n = 1$, we have $p_1(1 - p_1) = 2a_1'(1 - 2a_1') \ge a_1' \ge a_1$, and we are done.

Proof of (ii). For (ii), we employ the same embedding, but extend it to infinite sequences $X_1^*, X_2^*, \ldots$ and $X_1, X_2, \ldots$. We will need the quantities $J_n(b) = E(\int |f_n - f(b,.)|)$ and $\bar{J}_n(b) = \sup_{m \geq n} J_m(b)/a_m$. Let $D_n = \{b: \bar{J}_n(b) > 1\}$. Since D_n decreases monotonically, there is a limit set D, and thus $\int_{D_n \cap [0,1]} dx \downarrow \int_{D \cap [0,1]} dx = \lambda(D)$ (λ is Lebesgue measure). Let D_n^c be the complement of D_n.

The introduction of the set D_n was merely needed to allow us to use Fatou's lemma:

$$\sup_b \limsup_{n\to\infty} \frac{J_n(b)}{a_n} \geq \sup_b \limsup_{n\to\infty} \left(\frac{J_n(b)}{a_n} \right) I_{D_n^c}(b)$$

$$\geq E\left(\limsup_{n\to\infty} \left(\frac{J_n(B)}{a_n} \right) I_{D_n^c}(B) \right)$$

$$\geq \limsup_{n\to\infty} E\left(\left(\frac{J_n(B)}{a_n} \right) I_{D_n^c}(B) \right). \tag{4}$$

For the individual terms in (4) we will use the inequalities (1) and (2), suitably modified. If we write $X = (X_1, X_2, \ldots)$, then the left-hand side of (2) after modification becomes

$$E\left(I_{[N_i=0]} a_n^{-1} I_{D_n^c}(B) \int_{C_i} |f_n - f(B,.)| \,|X \right)$$

$$\geq I_{[N_i=0]} (2a_n)^{-1} E\left(I_{[J_n(B') \leq 1]} \int_{C_i} |f_n - f(B',.)| \right.$$

$$\left. + I_{[\bar{J}_n(B'') < 1]} \int_{C_i} |f_n - f(B'',.)| \,|X \right)$$

$$\geq I_{[N_i=0]} (2a_n)^{-1} E\left(I_{[\max(\bar{J}_n(B'), \bar{J}_n(B'')) \leq 1]} \int_{C_i} |f(B',.) - f(B'',.)| \,|X \right)$$

$$\geq I_{[N_i=0]} \frac{p_i}{a_n} E\left(I_{[\max(\bar{J}_n(B'), \bar{J}_n(B'')) \leq 1]} |X \right)$$

and the expected value of the last expression is

$$\begin{aligned} & p_i a_n^{-1} P\big(N_i = 0, \max\big(\bar{J}_n(B'), \bar{J}_n(B'')\big) \le 1\big) \\ & \quad \ge p_i a_n^{-1}\big(P(N_i = 0) - P\big(N_i = 0, \bar{J}_n(B') > 1\big) \\ & \qquad - P\big(N_i = 0, \bar{J}_n(B'') > 1\big)\big) \\ & \quad \ge p_i a_n^{-1}\big(P(N_i = 0) - 2P\big(N_i = 0, \bar{J}_n(B) > 1\big) \\ & \qquad - 2P\big(N_i = 0, \bar{J}_n(B) > 1\big)\big) \\ & \quad = p_i a_n^{-1}\big(P(N_i = 0) - 4P(N_i = 0, B \in D_n)\big). \end{aligned} \tag{5}$$

Let $A_n = [k_{n-1}, \infty)$, where k_n is as defined in the proof of Lemma 1. Define also

$$Z_n = \sum_{i \in A_n} p_i I_{[N_i = 0]}.$$

We have shown that

$$\begin{aligned} E\left(\frac{J_n(B)}{a_n} I_{D_n^c}(B)\right) & \ge E\left(\frac{Z_n}{a_n}\right) - 4E\left(\frac{Z_n}{a_n} I_{D_n}(B)\right) \\ & \ge E\left(\frac{Z_n}{a_n}\right) - 4\left(E\left(\frac{Z_n^2}{a_n^2}\right) P(B \in D_n)\right)^{1/2} \end{aligned}$$

(by Schwarz's inequality).

Now, we choose the probability vector $(p_1, p_2, \ldots)$ as dictated by the construction of Lemma 1. Then

$$E\left(\frac{Z_n}{a_n}\right) = a_n^{-1} \sum_{i \ge k_{n-1}} p_i (1 - p_i)^n \not\to 0.$$

Furthermore,

$$\begin{aligned} E(Z_n^2) &= \sum_{i \in A_n} p_i^2 (1 - p_i)^n + \sum_{i \ne j;\, i, j \in A_n} p_i p_j (1 - p_i - p_j)^n \\ &\le 2 \sum_{i \in A_n} p_i^2 (1 - p_i)^{2n} + \sum_{i \ne j;\, i, j \in A_n} p_i (1 - p_i)^n p_j (1 - p_j)^n \\ &\le 2E^2(Z_n), \end{aligned}$$

where we used the fact that on A_n, $(1-p_i)^n \geq \frac{1}{2}$, and that for all i, j: $1-p_i-p_j \leq (1-p_i)(1-p_j)$. Combining the last two facts and (5), (4), and $\lim_{n\to\infty} P(B \in D_n) = \lambda(D)$, we conclude that

$$\sup_b \limsup_{n\to\infty} \frac{J_n(b)}{a_n} \geq \left(1 - 4\sqrt{2\lambda(D)}\right) \limsup_{n\to\infty} E\left(\frac{Z_n}{a_n}\right).$$

We always have

$$\sup_b \limsup_{n\to\infty} \frac{J_n(b)}{a_n} \geq \lambda(D)$$

by definition of D.

Because $\lambda(D) > 0$ certainly implies that for some b, $\limsup_{n\to\infty} J_n(b)/a_n > 0$, and because $\lambda(D) = 0$ implies the same thing by the former inequality, we can conclude that there exists some b such that

$$\limsup_{n\to\infty} \frac{J_n(b)}{a_n} > 0.$$

But since we can always replace a_n by $\sqrt{a_n}$ in our choice of a probability vector $(p_1, p_2, \ldots)$, we see that (ii) follows from this result for families G and $H(g)$.

Class G_∞. We will change the definition of family 2 very slightly. Consider the density $g(x) = C\exp(-1/x(1-x))$ on $[0,1]$. Then, let $C_1, C_2, \ldots$ be the intervals $[0, p_1], [p_1, p_1+p_2], \ldots$. Dichotomize each interval into two intervals of equal length, and call the left interval A_i and the right interval A_i'. Next, define

$$f(b,x) = \sum_{i=1}^{\infty} 2(b_i g_i + (1-b_i) g_i'),$$

where g_i is $g(x/(p_i/2))$ translated to $p_1 + \cdots + p_{i-1}$, and g_i' is g_i translated $p_i/2$ to the right. Thus, g_i vanishes outside A_i and g_i' vanishes outside A_i'. Furthermore, family 2 belongs to G_∞. To see this, we need only look at the maximal value of f, that is, twice the maximal value of g: $2C\exp(-4)$.

Now,

$$\int_0^1 \exp\left(-\frac{1}{x(1-x)}\right) = 2\int_0^{1/2} \exp\left(-\frac{4}{1-4y^2}\right) dy$$

$$= 2e^{-4}\int_0^{1/2} \exp\left(-\frac{16y^2}{1-4y^2}\right) dy$$

$$\geq 2e^{-4}\int_0^{1/\sqrt{20}} (1-20y^2)\, dy,$$

where we used the inequality $e^{-u} \geq 1-u$, and the fact that $1-4y^2 \leq 1$. But the lower bound is $\frac{4}{3}e^{-4}/\sqrt{20}$. Therefore, $2Ce^{-4} \leq \frac{3}{2}\sqrt{20} = 3\sqrt{5}$. It is easy to see that for any $\delta > 0$ in the definition of G_∞, we can choose an appropriate g, flat on $[\varepsilon, 1-\varepsilon]$, and very smooth on $[0, \varepsilon]$ and $[1-\varepsilon, 1]$. It is simple to establish that the proof of Theorem 1 carries over to this family *without change*.

The Class U. We will consider a subclass of U parametrized by b and $(p_1, p_2, \ldots)$ as before. The vector of p_i's is nonincreasing in i and sums to $\frac{1}{2}$. We first determine intervals by defining knots

$$x_0 = 0, \quad x_1 = 1, \quad x_{i+1} - x_i = 2^i p_i, \qquad i \geq 1.$$

The density $f(b,.)$ is constant on $[x_i, x_{i+1})$ and 0 elsewhere. For $x \in [x_i, x_{i+1})$, its value is determined as follows:

$$\text{if } i = 0\text{: } f(x) = \frac{1}{2} + \sum_{\substack{i \geq 1 \\ b_i = 1}} \frac{p_i}{2};$$

$$\text{if } i \neq 0, b_i = 0\text{: } f(x) = 2^{-i};$$

$$\text{if } i \neq 0, b_i = 1\text{: } f(x) = 2^{-(i+1)}.$$

Note that f is unimodal on $[0, \infty)$, and that it is a density in view of

$$\int f = (x_1 - x_0)\left(\frac{1}{2} + \sum_{\substack{i \geq 1 \\ b_i = 1}} \frac{p_i}{2}\right) + \sum_{\substack{i \geq 1 \\ b_i = 1}} (x_{i+1} - x_i)2^{-i-1}$$

$$+ \sum_{\substack{i \geq 1 \\ b_i = 0}} (x_{i+1} - x_i)2^{-i}$$

$$= \frac{1}{2} + \sum_{\substack{i \geq 1 \\ b_i = 1}} \frac{p_i}{2} + \sum_{\substack{i \geq 1 \\ b_i = 1}} \frac{p_i}{2} + \sum_{\substack{i \geq 1 \\ b_i = 0}} p_i = \frac{1}{2} + \sum_{i=1}^{\infty} p_i = 1.$$

The embedding is slightly more difficult now. We start with independent random variables $B, X_1^*, \ldots, X_n^*, Y_1, \ldots, Y_n$, where B is as before, the X_i^*'s have density $f(0,.)$ (recall that $f(0,x) = \frac{1}{2}$ on $[x_0, x_1)$; and $f(0,x) = 2^{-i}$ on $[x_i, x_{i+1}), i \geq 1$), and the Y_i's are Bernoulli random variables with parameter $\frac{1}{2}$. We define our sample $X_1, \ldots, X_n$ from $f(b,.)$ by

$$X_j = X_j^* \quad \text{if } X_j^* \in [x_0, x_1)$$
$$\text{or if } X_j^* \in [x_i, x_{i+1}), \quad i \geq 1, \text{ and } b_i = 0$$
$$\text{or if } X_j^* \in [x_i, x_{i+1}), \quad i \geq 1, \text{ and } b_i = 1, Y_j = 0$$

$$X_j = \frac{(X_j^* - x_i)}{(x_{i+1} - x_i)} \quad \text{if } X_j^* \in [x_i, x_{i+1}), i \geq 1, b_i = 1, Y_j = 1.$$

Thus, in the last case, we replace X_j^* by a uniform $[x_0, x_1)$ random variable. It is easy to verify that the X_j's have density $f(b,.)$.

Check that (1) remains valid after replacing C_i by $[x_i, x_{i+1})$. Then, because on $N_i = 0$ $(i \geq 1)$, where N_i is the number of X_j's in $[x_i, x_{i+1})$, B_i and $X_1, \ldots, X_n$ are conditionally independent, we can derive (2) again for all $i \geq 1$ with the following modifications: condition on $X_1^*, Y_1, \ldots, X_n^*, Y_n$, and replace p_i by $p_i/4$ on the last line. Thus, we can conclude that (3) holds if the right-hand side is divided by 4.

This lower bound depends upon n and the vector of p_i's. As we have shown above, we can find a nonincreasing sequence of p_i's with sum equal to $\frac{1}{2}$, such that $\frac{1}{4}\sum_{i=1}^{\infty}(p_i(1-p_i)^n/a_n) \to \infty$ as $n \to \infty$. Also, the supremum of the lower bound over all such vectors of p_i's is $\frac{1}{8}$. This concludes the proof of (i) for this class of densities. The proof of (ii) poses no major problems either.

The Class U_∞. Consider the family of piecewise rectangular unimodal densities constructed for U, and construct a new family taken from U_∞ consisting of similar piecewise "almost" rectangular unimodal densities: on each interval $[x_i, x_{i+1})$, define $f(b,.)$ as before except near x_i and x_{i+1}, where a continuity correction is made. Since U can be considered as a limiting case of U_∞, it is not hard to see that all the previous results remain valid here.

5. PROOF OF THEOREMS 2 AND 11

Note that Theorem 2 is but a special case of Theorem 11 (take $p = 1$). Consider the family 1 of the proof of Theorem 1 once again. It is not hard

to check that

$$\int f^p(b,.) = \left(\sum p_i^p\right)\int g^p,$$

and that

$$D_{sp}(f) = D_{sp}(g)\left(\sum p_i^p\right)^{1/(2s+1)}\left(\sum p_i^{p/2}\right)^{2s/(2s+1)}.$$

Now, rederive (1) and (2) up to the second-to-last expression of (2), in which the factor $\frac{1}{2}$ should be replaced by 2^{-p}. The integral over C_i of $\int|f(B',.) - f(B'',.)|^p$ is $2p_i^p\int g^p$, and this gives

$$\sup_b E\left(\int|f_n - f(b,.)|^p\right) \geq 2^{-(p-1)}\int g^p \sum p_i^p(1-p_i)^n.$$

Now, take $p_i = 1/M$ for $i = 1, 2, \ldots, M$, and 0 elsewhere. Then

$$\sup_b \frac{E\left(\int|f_n - f(b,.)|^p\right)}{D_{sp}(f)} \geq \frac{\int g^p}{D_{sp}(g)} 2^{-(p-1)}\left(\frac{1}{M}\right)^{ps/(2s+1)}\left(1 - \frac{1}{M}\right)^n.$$

Take $M \sim n/(ps/(2s+1))$, note that $\int g^p \geq 1$, and let $n \to \infty$.

6. PROOF OF THEOREM 3

We will use a randomization technique as in the proof of Theorem 1 (i) (see (1)). A family of densities strictly contained in $F_{s,r}$ is constructed around a central density g. This central density in turn is obtained as $g_0 * g_1$ (a convolution) where g_0 is a density with support in $[-\frac{1}{4}, \frac{1}{4}]$ and continuous $(s-1)$st derivative, and g_1 is the uniform density on $[-\frac{1}{2}, \frac{1}{2}]$. Thus,

$$\begin{aligned}(g_0 * g_1(x))^{(s)} &= \int_{x-1/2}^{x+1/2} g_0^{(s)}(y)\,dy \\ &= g_0^{(s-1)}\left(x + \tfrac{1}{2}\right) - g_0^{(s-1)}\left(x - \tfrac{1}{2}\right),\end{aligned}$$

and

$$\begin{aligned}\int|(g_0 * g_1)^{(s)}| &= \int|g_0^{(s-1)}\left(x + \tfrac{1}{2}\right) - g_0^{(s-1)}\left(x - \tfrac{1}{2}\right)|\,dx \\ &= 2\int|g_0^{(s-1)}(x)|\,dx.\end{aligned}$$

To randomize, we will consider a uniform $[0,1]$ random variable B with binary expansion $B = 0.B_1B_2\ldots$ in which all the bits are $+1$ or -1 (i.e., all occurrences of 0 are changed to -1). Individual realizations of B will be denoted by b. We will also need h_j, u_j, $j \geq 1$, sequences of real numbers with the property that $6\Sigma h_j \leq 1$, $0 \leq u_j \leq 1$. Now, find real numbers x_j, $j \geq 1$, such that the sets $A_j = [x_j - \frac{3}{2}h_j, x_j + \frac{3}{2}h_j]$ are disjoint and contained in $\{x: g(x) = 1\}$. (This can be done because $6\Sigma h_j \leq 1$.)

For fixed $b \in [0,1)$, we define the density $f(b,x)$ in our family as follows:

$$f(b,x) = g(x)\left(1 + b_j u_j\left(g\left(\frac{x - x_j}{h_j} + \frac{3}{4}\right) - g\left(\frac{x - x_j}{h_j} - \frac{3}{4}\right)\right)\right),$$

$$x \in A_j,$$

and $f(b,x) = g(x)$ when x does not belong to any A_j. One can easily check that $f \geq 0$ because $|b_j u_j| \leq 1$ for all j and $g \leq 1$. Also, because of the construction of g, the integral of f is equal to the integral of g.

Clearly, suppressing the dependence of f_n upon a sample with density $f(b,.)$ and integration with respect to dx, we have

$$\sup_{0 \leq b < 1} E\left(\int |f_n(x) - f(b,x)|\right) \geq \int_0^1 E\left(\int |f_n(x) - f(b,x)|\right) db$$

$$\geq \sum_j \int_0^1 E\left(\int_{A_j} |f_n(x) - f(b,x)|\right) db. \quad (6)$$

Consider now an individual term in (6), and let us fix $b_1, b_2, \ldots, b_{j-1}, b_{j+1}, \ldots$. Let b^+, b^- be the corresponding numbers with the given binary expansion into b_i's with $b_j = +1$ and $b_j = -1$, respectively. Thus, averaging the jth term in (6) with respect to the jth bit b_j and with respect to "E" only gives

$$\frac{1}{2}\left(E\left(\int_{A_j} |f_n(x) - f(b^+,x)|\right) + E\left(\int_{A_j} |f_n(x) - f(b^-,x)|\right)\right).$$

Introduce the notation $y = (z_1, \ldots, z_n) \in R^n$; $u^+ = u^+(y) = \int_{A_j} |f_n(x;y) - f(b^+,x)|$; $u^- = \int_{A_j} |f_n(x;y) - f(b^-,x)|$; $f^+(y) = \prod_{i=1}^n f(b^+, z_i)$; and $f^-(y)$

$= \prod_{i=1}^{n} f(b^-, z_i)$. Then, the last expression is bounded from below by

$$\frac{1}{2}\left(\int u^+(y)f^+(y)\,dy + \int u^-(y)f^-(y)\,dy\right)$$

$$\geq \frac{1}{2}\int (u^+(y) + u^-(y))\min(f^+(y), f^-(y))\,dy$$

$$\geq \frac{1}{2}u\int \min(f^+, f^-)$$

$$\geq \frac{1}{4}u\exp\left(-\int f^+\log\left(\frac{f^+}{f}\right)\right), \tag{7}$$

where $u = \int_{A_j} |f(b^+, x) - f(b^-, x)|$ does not depend upon the b_i's, $i \neq j$. The last inequality in (7) is proved in Theorem 8.2. The exponent in (7) is equal to $n\int f(b^+, x)\log(f^-(b, x)/f^+(b, x))$. Also, because for $|z| \leq 1$,

$$\left|(1+z)\log\left(\frac{1-z}{1+z}\right) - 2z\right| \leq \frac{2z^2}{1-|z|},$$

we have, if we write f as $g(1 + b_j g_1)$ on A_j,

$$\int f^+\log\left(\frac{f^-}{f^+}\right) = n\int_{A_j} g(1+g_1)\log\left(\frac{1-g_1}{1+g_1}\right)$$

$$\geq \int_{A_j} 2gg_1 - \int_{A_j} \frac{2gg_1^2}{1-|g_1|}$$

$$\geq 0 - \frac{4nu_j^2 h_j}{1-u_j}.$$

Because $u = 2\int |g_1| \geq 2u_j h_j$, we can combine all the previous bounds into one, and note that (6) is at least equal to

$$\sum_j \frac{1}{2}u_j h_j \exp\left(-\frac{4nu_j^2 h_j}{1-u_j}\right). \tag{8}$$

Let us take $u_j = u$, $h_j = h$, $1 \leq j \leq N$, such that $6hN = 1$, $u/h^s = a > 0$. With this choice, we can compute $D_s(f)$ for each b. In particular, we always

have

$$\sqrt{1-u}\int\sqrt{g} \le \int\sqrt{f} \le \sqrt{1+u}\int\sqrt{g},$$

$$\int|f^{(s)}| = (1+2Nuh^{1-s})\int|g^{(s)}| = \left(1+\frac{a}{3}\right)\int|g^{(s)}|,$$

and, thus, because we will let $N \to \infty$, $u \to 0$,

$$D_s(f) \to D_s(g)\left(1+\frac{a}{3}\right)^{1/(2s+1)}$$

The lower bound (8), rewritten as $(u/12)\exp(-4nu^2(u/a)^{1/s}/(1-u))$ is approximately minimized by taking $u = ah^s$, $h^{2s+1} = (s/(2s+1))/4na^2$, and the corresponding value of the lower bound is

$$\frac{1}{12}\left(\frac{s}{4ne(2s+1)}\right)^{s/(2s+1)} a^{1/(2s+1)} \ge \frac{1}{12}e^{-4}a^{1/(2s+1)}n^{-s/(2s+1)}.$$

To make all densities $f(b,.)$ asymptotically belong to $F_{s,r}$, we must take $(1+a/3)^{1/(2s+1)}D_s(g) \le r$. Thus, let us pick $a = 3((r/D_s(g))^{2s+1} - 1)$ (this is certainly positive for $r > D_s(g)$). Then, because $3^{1/(2s+1)}/12 \ge 2^{-4}$ and $((r/D_s(g))^{2s+1} - 1)^{1/(2s+1)} \ge r/D_s(g) - 1$, the lower bound is

$$(2e)^{-4}\frac{r/D_s(g)-1}{n^{s/(2s+1)}},$$

valid for all $r > D_s(g)$ and all n large enough. This concludes the proof of Theorem 3, if we can establish that at least for some $g = g_0 * g_1$, we have $D_s(g) \le (\frac{1}{4}9^s(s+1)!)^{1/(2s+1)}$. To see this, note that

$$\int\sqrt{g} \le \frac{6}{4} = \frac{3}{2} \quad (\text{because } g_0 * g_1 \le 1 \text{ on } [-\tfrac{3}{4}, \tfrac{3}{4}]).$$

Also, if we take $g_0(x) = 2(s+1)(1-4|x|)^s$, $4|x| \le 1$, then

$$\begin{aligned}\int|g^{(s)}| &= 2\int|g_0^{(s-1)}| \\ &= 2.4^{s-1}2(s+1)2\int_0^{1/4}(1-4x)\,dx\,s! \\ &= 4^{s-1}(s+1)!.\end{aligned}$$

Thus, for this particular choice of g,

$$D_s(g) \le \left(\left(\tfrac{3}{2}\right)^{2s} 4^{s-1}(s+1)!\right)^{1/(2s+1)},$$

which was to be shown.

7. PROOF OF THEOREM 4

Let $\tilde{f}_n$ be a density estimate. Then the following estimate is strictly better:

$$f_n(x) = \tilde{p}g(x) + (1-\tilde{p})g(x+2),$$

where $\tilde{p} = \int_{[0,1]} \tilde{f}_n$, because

$$\begin{aligned}\int |f_n - f| &= \int_{[0,1]} g|\tilde{p} - p| + \int_{[0,1]} g|(1-\tilde{p}) - (1-p)| \\ &= \left|\int_{[0,1]} \tilde{f}_n - f\right| + \left|\int_{([0,1])^c} \tilde{f}_n - f\right| \\ &\le \int |\tilde{f}_n - f|.\end{aligned}$$

In other words, when we derive lower bounds, we can assume without loss of generality that $f_n(x) = \tilde{p}g(x) + (1-\tilde{p})g(x+2)$, where $\tilde{p} = \tilde{p}(X_1, \ldots, X_n)$ is a Borel measurable function of its arguments taking values in $[0,1]$. Now, let us define $N = \sum_{i=1}^n I_{[X_i \in [0,1]]}$. Then,

$$E\left(\int |f_n - f|\right) = 2E(|\tilde{p} - p|) \ge 2E(|E(\tilde{p}|N) - p|) = 2E(|\psi(N) - p|)$$

for some $[0,1]$-valued Borel measurable function ψ. Here we used the conditional version of Jensen's inequality, and the fact that N is a sufficient statistic for p.

Let us write E_p for the expected value with respect to the distribution of $(X_1, \ldots, X_n)$ when the mixing parameter in f is p. Let N_p and N_q be binomial (n, p) and (n, q) random variables, respectively, where $0 \le p \le q \le 1$. The fundamental inequality underlying the remainder of the proof is

$$\frac{1}{2}\left(E_p\left(\int |f_n - f|\right) + E_q\left(\int |f_n - f|\right)\right) \ge |p - q| P(N_q \le np).$$

The proof of this inequality is simple:

$$\begin{aligned}
&\frac{1}{2}\left(E_p\left(\int|f_n - f|\right) + E_q\left(\int|f_n - f|\right)\right) \\
&\quad \geq E\left(|\psi(N_p) - p|\right) + E\left(|\psi(N_q) - q|\right) \\
&\quad \geq \sum_{j \leq np} \left(|\psi(j) - p|P(N_p = j) + |\psi(j) - q|P(N_q = j)\right) \\
&\quad \geq \sum_{j \leq np} \left(|\psi(j) - p| + |\psi(j) - q|\right)P(N_q = j) \\
&\quad \geq |p - q|P(N_q \leq np).
\end{aligned}$$

Part (i) of Theorem 4 follows directly from our fundamental inequality by a randomization argument. Let p be a random variable with probability measure μ on $[0, 1]$. Then,

$$\sup_p E_p\left(\int|f_n - f|\right) \geq \int E_p\left(\int|f_n - f|\right)\mu(dp).$$

Let μ put half its mass at $a = \frac{1}{2}$, and half its mass at $b = \frac{1}{2} + c/\sqrt{n}$ for some constant c. From the fundamental inequality we note that

$$\sup_p E_p\left(\int|f_n - f|\right)\sqrt{n} \geq cP(N_b \leq na) \sim c\Phi(-2c),$$

where Φ is the normal $(0, 1)$ distribution function. The last step follows from the central limit theorem. The lower bound $c\Phi(-2c)$ is maximal for $2c = 0.7517915241\cdots$ and takes the value of $0.0849856\cdots$. If a bound is needed for particular values of n, we can follow many routes: for example, assume that $c = 19/100$, $n \geq 4$ (these choices will be convenient). Note that $n(1 - b) \leq \lfloor na \rfloor \leq nb$. Slud (1977) has shown that

$$cP(N_b \leq na) \geq c\Phi\left(\frac{\lfloor na \rfloor - nb}{\sqrt{nb(1-b)}}\right).$$

For n even, this is at least equal to $c\Phi(-c/\sqrt{\frac{1}{4} - c/2\sqrt{n}}) \geq c\Phi(-c/\sqrt{\frac{1}{4} - c/4}) = c\Phi(-2c/\sqrt{1-c})$. For n odd, a lower bound is provided by $c\Phi((-2c - 1/\sqrt{n})/\sqrt{1-c}) \geq c\Phi((-2c - \frac{1}{2})/\sqrt{1-c})$. The odd lower bound, which is smaller than the even lower bound, is $\frac{19}{100}\Phi(-\frac{44}{45}) \geq \frac{19}{100}\Phi(-1) = 0.30153\cdots$. This concludes the proof of (i).

For the proof of (ii), we introduce the notation $J_n(p) = E_p(\int |f_n - f|)\sqrt{n}$, and let p be a random variable with uniform probability measure μ on $[\frac{1}{2} - \varepsilon, \frac{1}{2} + \varepsilon]$ for some small $\varepsilon > 0$. Also, define $q = c/\sqrt{n}$, and choose c as in the proof of part (i), that is, $2c = 0.7517915241\cdots$. Let $a > 0$ be an arbitrary constant. By Fatou's lemma,

$$\sup_p \limsup_{n\to\infty} J_n(p) \geq E\Big(\limsup_{n\to\infty} J_n(p)\Big) \geq E\Big(\limsup_{n\to\infty} \min(J_n(p), a)\Big)$$

$$\geq \limsup_{n\to\infty} E\big(\min(J_n(p), a)\big)$$

$$= \limsup_{n\to\infty} \tfrac{1}{2}\big(E\big(\min(J_n(p), a) + E\big(\min(J_n(p \oplus q), a)\big)\big)$$

$$\geq \limsup_{n\to\infty} \tfrac{1}{2}E\big(\min\big(\tfrac{1}{2}(J_n(p) + J_n(p \oplus q)), a\big)\big)$$

$$\geq \limsup_{n\to\infty} \tfrac{1}{2}E\big(\min\big(cP(N_{p+q} \leq np|p), a\big)I_{p+q\leq 1/2+\varepsilon}\big),$$

where $p \oplus q$ is defined as $p + q$ when the sum is less than $\frac{1}{2} + \varepsilon$, and as $p + q + 2k\varepsilon$ for some other k integer. The integer k is chosen such that $p + q + 2k\varepsilon$ belongs to $[\frac{1}{2} - \varepsilon, \frac{1}{2} + \varepsilon]$. The symbol N_{p+q} is used for a binomial $(n, p + q)$ random variable. If we take $a = c$, then the last term in this chain of inequalities is equal to

$$\limsup_{n\to\infty} \tfrac{1}{2}E\big(cP(N_{p+q} \leq np|p)I_{p+q\leq 1/2+\varepsilon}\big).$$

Now, for fixed p such that $p \leq \frac{1}{2} + \varepsilon$, $cP(N_{p+q} \leq np|p) \to c\Phi(-c/\sqrt{p(1-p)})$, by the central limit theorem. Thus, by the dominated convergence theorem, the limit supremum is at least $\frac{1}{2}cE(\Phi(-c/\sqrt{p(1-p)}))$. Since $|p - \frac{1}{2}| \leq \varepsilon$, and since we can choose ε, this limit can be made arbitrarily close to $\frac{1}{2}c\Phi(-2c) = 0.0424928\cdots$. This concludes the proof of Theorem 4.

8. PROOF OF THEOREMS 5, 6, 7, 8, AND 9

Proof of Theorem 5. We proceed again by a randomization argument, using a uniform distribution over all 2^r possible values for $b = (b_1, \ldots, b_r)$. We will write Σ_b for the sum over all these values, b_{i+} and b_{i-} for r-vectors equal to b in all their components except possibly the ith: the ith compo-

nent of b_{i+} is always $+1$, and the ith component of b_{i-} is always -1. We will write $\mathbf{X}_n$ for $(X_1, \ldots, X_n) \in R^n$, and $\mathbf{x}_n$ for $(x_1, \ldots, x_n) \in R^n$. Finally, all products $\prod$ are over $j = 1, n$ as in $\prod f(b, x_j)$. The density estimate is $f_n(x), x \in R$, but it will be more convenient to write $f_n(x, \mathbf{X}_n)$ to make the dependence upon the data explicit, where $\mathbf{X}_n$ is a sample of n i.i.d. random variables with density $f(b, x)$. We have

$$\sup_b E\left(\int |f_n - f(b,.)|\right)$$

$$\geq 2^{-r} \sum_b \int \int |f_n(x, \mathbf{x}_n) - f(b,x)|\, dx \cdot \prod f(b, x_j)\, d\mathbf{x}_n$$

$$\geq 2^{-r} \sum_b \int \sum_{i=1}^{r} \int_{A_i} |f_n(x, \mathbf{x}_n) - f(b,x)|\, dx \cdot \prod f(b, x_j)\, d\mathbf{x}_n$$

$$= 2^{-r} \sum_b \int \sum_{i=1}^{r} \frac{1}{2} \left(\int_{A_i} |f_n(x, \mathbf{x}_n) - f(b_{i+}, x)|\, dx \cdot \prod f(b_{i+}, x_j) \right.$$

$$\left. + \int_{A_i} |f_n(x, \mathbf{x}_n) - f(b_{i-}, x)|\, dx \cdot \prod f(b_{i-}, x_j) \right) d\mathbf{x}_n$$

$$\geq 2^{-r} \sum_b \int \sum_{i=1}^{r} \frac{\alpha}{2} \operatorname{Min}\left(\prod f(b_{i+}, x_j), \prod f(b_{i-}, x_j)\right) d\mathbf{x}_n$$

$$\geq \frac{r\alpha}{2} \inf_{b,i} \int \operatorname{Min}\left(\prod f(b_{i+}, x_j), \prod f(b_{i-}, x_j)\right) d\mathbf{x}_n.$$

Now, if f and g are two densities, we know that $\int \operatorname{Min}(f, g) \geq \frac{1}{2}\left(\int \sqrt{fg}\right)^2$ (Theorem 8.5) and that $\int \operatorname{Min}(f, g) = 1 - \frac{1}{2}\int |f - g| \geq 1 - \left(\int (\sqrt{f} - \sqrt{g})^2\right)^{1/2}$ (Theorem 8.4). But

$$\int \sqrt{\prod f(b_{i+}, x_j) \prod f(b_{i-}, x_j)}\, d\mathbf{x}_n = \prod \int \sqrt{f(b_{i+}, x_j) f(b_{i-}, x_j)}\, dx_j \geq \beta^n$$

and the first part of Theorem 5 is proved.

For the second part, note that $\alpha = 2\int_A |g|$ and that $\gamma \leq \int_A g^2$. This follows from the observation that for any b and i,

$$1 - \int \sqrt{f(b_{i+}, x) f(b_{i-}, x)} = \int_{A + y_i} \left(f_0 - \sqrt{f_0 + g}\sqrt{f_0 - g}\right)$$

$$= \int_{A + y_i} \left(1 - \sqrt{1 - g^2}\right) \leq \int_A g^2.$$

REMARK. The first part of the proof of Theorem 5 essentially coincides with the chain of inequalities (6), (7) in the proof of Theorem 3, due to Bretagnolle and Huber (1979).

Proof of Theorem 6. We will first find a nonnegative function g_0 satisfying the $W(s, \alpha, C)$ condition and

$$\int_0^1 |g_0| \geq \gamma_1 C, \quad \int_0^1 g_0^2 \leq \gamma_2 C^2, \quad \sup g_0 \leq \gamma_3 C.$$

After having done this, we proceed as follows to construct a family as in Theorem 5. Take a constant $a_0 > 1$. Define $A = [0, 1/a_0 r)$, and

$$g(x) = \begin{cases} \dfrac{g_0(2a_0 r x)}{(2a_0 r)^{s+\alpha}}, & 0 \leq x \leq \dfrac{1}{2a_0 r}, \\ \dfrac{-g_0(2a_0 r(x - 1/(2a_0 r)))}{(2a_0 r)^{s+\alpha}}, & \dfrac{1}{2a_0 r} \leq x < \dfrac{1}{a_0 r}. \end{cases}$$

Let y_i be $\frac{1}{2}(1 - 1/a_0) + i/a_0 r$, $i = 0, 1, \ldots, r - 1$. Thus, $\bigcup_i A + y_i = [\frac{1}{2}(1 - 1/a_0), \frac{1}{2}(1 + 1/a_0))$, that is, we have cut the latter interval into r equal pieces of length $1/a_0 r$ each. The function f_0 of Theorem 5 is 1 on this big central interval, and must be defined carefully outside it. In particular, we should make sure that $f_0 = 0$ outside $[0, 1]$, that $\int_0^1 f_0 = 1$, that $f_0 \geq 0$, that $f_0^{(i)}(0) = 0$ for $i = 0, 1, \ldots, s - 1$, that $f_0(\frac{1}{2}(1 - 1/a_0)) = 1$, and that $f_0^{(i)}(\frac{1}{2}(1 - 1/a_0)) = 0$, $i = 1, \ldots, s - 1$. Furthermore, f_0 should satisfy the $W(s, \alpha, C)$ condition, and the interval $[\frac{1}{2}(1 + 1/a_0), 1]$ should be treated symmetrically. This can always be done merely by choosing C larger than a threshold $c_1(s, \alpha, a_0)$. Later we will take $a_0 = 2$ (for no particular reason except convenience), and this allows us to introduce the condition $C \geq c_1(s, \alpha)$. Let us quickly verify that the family of Theorem 5 is indeed contained in $W(s, \alpha, C)$. First,

$$g^{(s)}(x) = \frac{g_0^{(s)}(2a_0 r x)}{(2a_0 r)^{\alpha}}, \qquad 0 \leq x \leq \frac{1}{2a_0 r},$$

and thus

$$|g^{(s)}(x) - g^{(s)}(y)| \leq (2a_0 r)^{-\alpha} C (2a_0 r)^{\alpha} |x - y|^{\alpha} = C|x - y|^{\alpha}.$$

Also,

$$\int_A |g| = 2\int_0^1 \frac{|g_0|}{(2a_0 r)^{s+\alpha+1}} \geq \frac{2\gamma_1 C}{(2a_0 r)^{s+\alpha+1}},$$

$$\int_A g^2 = 2\int_0^1 \frac{g_0^2}{(2a_0 r)^{2(s+\alpha)+1}} \leq \frac{2\gamma_2 C^2}{(2a_0 r)^{2(s+\alpha)+1}},$$

and

$$\sup_A |g| = \sup_{[0,1]} \frac{|g_0|}{(2a_0 r)^{s+\alpha}} \leq \frac{\gamma_3 C}{(2a_0 r)^{s+\alpha}}.$$

The lower bound of Theorem 5 now reads

$$\frac{1}{2} r \int_A |g| \geq \gamma_1 C r (2a_0 r)^{-s-\alpha-1} = \gamma_1 C (2a_0)^{-s-\alpha-1} r^{-(s+\alpha)}$$

subject to the conditions

$$n\int_A g^2 \leq \frac{2n\gamma_2 C^2}{(2a_0 r)^{2(s+\alpha)+1}} \leq \frac{1}{8}$$

and

$$\frac{\gamma_3 C}{(2a_0 r)^{s+\alpha}} \leq 1.$$

The first one of these side conditions is used to determine r; to obtain the best possible result, we should take r equal to the ceiling function (i.e., next integer greater than or equal to) of

$$(2a_0)^{-1}(16\gamma_2 n C^2)^{1/(2(s+\alpha)+1)}.$$

The last condition now becomes a lower bound for n. For example, it is satisfied if

$$(16\gamma_2 n C^2)^{(s+\alpha)/(2(s+\alpha)+1)} \geq \gamma_3 C,$$

which is equivalent to the condition $n \geq Cc_2$. Resubstituting our value for r

in the lower bound, with $a_0 = 2$, gives us the asymptotic result mentioned in Theorem 6. By using the fact that ceiling $(y) \le y + 1$, we obtain the lower bound

$$\frac{\gamma_1 C/4}{\left(\left(16\gamma_2 n C^2\right)^{1/(2s+2\alpha+1)} + 4\right)^{s+\alpha}},$$

valid for all $n \ge Cc_2$.

This leaves us with the task of finding our original function g_0. We consider deliberately a suboptimal but practical g_0, namely,

$$g_0(x) = C\gamma_0(x(1-x))^{s+\alpha}.$$

We verify without effort that $\int_0^1 |g_0| = C\gamma_1$, $\int_0^1 g_0^2 = C^2\gamma_2$, and $\sup g_0 = C\gamma_3$. Thus, Theorem 6 is proved if we can prove that g_0 satisfies the $W(s, \alpha, C)$ condition. Using the notation

$$\binom{u}{j} = 1 \quad (\text{for } j = 0) \quad \text{and} \quad \frac{u(u-1)\cdots(u-j+1)}{j!} \quad (\text{for } j > 0), \quad u \in R,$$

and the binomial expansion theorem, we see that

$$\frac{g_0(x)}{C\gamma_0} = \sum_{j=0}^{\infty} (-1)^j \binom{s+\alpha}{j} x^{s+\alpha+j}.$$

The sth derivative of $g_0(x)/(C\gamma_0)$ is $\sum_{j=0}^{\infty}(-1)^j h_j(x)$, where

$$h_j(x) = \binom{s+\alpha}{j}(s+\alpha+j)\cdots(1+\alpha+j)x^{\alpha+j}.$$

Now, for $0 \le x \le y \le \frac{1}{2}$,

$$|h_j(x) - h_j(y)| = \binom{s+\alpha}{j}(s+\alpha+j)\cdots(1+\alpha+j)|x^{\alpha+j} - y^{\alpha+j}|$$

and

$$|y^{\alpha+j} - x^{\alpha+j}| \le \begin{cases} |y-x|, & j = 0, \\ (\alpha+j)y^{\alpha+j-1}|y-x|, & j > 0. \end{cases}$$

Combining all these estimates in a super-upper bound, we obtain

$$\frac{|g_0^{(s)}(x) - g_0^{(s)}(y)|}{C\gamma_0}$$

$$\leq (s+\alpha)\cdots(1+\alpha)|y-x| + \sum_{j=1}^{\infty}\binom{s+\alpha}{j}(s+\alpha+j)\cdots(\alpha+j)\,y^{\alpha+j-1}|y-x|$$

$$\leq |y-x|\left((s+\alpha)\cdots(1+\alpha) + \sum_{j=1}^{\infty}\frac{(s+\alpha+j)^{s+1}(s+\alpha)^j}{2^{\alpha+j-1}j!}\right)$$

$$\leq |y-x|\left((s+\alpha)\cdots(1+\alpha) + \sum_{j=1}^{\infty}(s+\alpha)^{s+1}2^{1-\alpha}\frac{((s+\alpha)e^{(s+1)/(s+\alpha)}/2)^j}{j!}\right)$$

$$\leq |y-x|(s+\alpha)^{s+1}2^{1-\alpha}\exp\left(\frac{s+\alpha}{2}\exp\left(\frac{s+1}{s+\alpha}\right)\right)$$

where we used the inequality $(1+u) \leq e^u$, $y \in R$. But the right-hand side is equal to $|y-x|/\gamma_0$, which was to be shown.

Proof of Theorem 7. In the first part for $W(0,1,C)$, we will illustrate how the inequality of Theorem 5 can be manipulated by optimization methods to get good bounds. We follow the proof of Theorem 6 first, and use the following function g_0 on $[0,1]$:

$$g_0(x) = \begin{cases} Cx, & 0 \leq x \leq \frac{1}{2}, \\ C(1-x), & \frac{1}{2} \leq x \leq 1. \end{cases}$$

Obviously, $\int_0^1 |g_0| = C/4$, $\int_0^1 g_0^2 = C^2/12$, and $\sup g_0 = C/2$. Thus, the constants γ_1, γ_2, and γ_3 in Theorem 6 can formally be replaced by $\frac{1}{4}$, $\frac{1}{12}$, and $\frac{1}{2}$. Define g from g_0 as in Theorem 6. We will let $a_0 > 1$ and $r \geq 1$ keep their meaning from Theorem 6 too. Thus, by Theorem 5, for C large enough,

$$\sup_{f \in W(0,1,C)} E\left(\int |f_n - f|\right) \geq r\int_A |g|\left(1 - \sqrt{2n\int_A g^2}\right)$$

$$= \frac{\gamma_1 C}{a_0}\left(\frac{1}{y} - \frac{L}{y^{5/2}}\right)$$

where $L = \sqrt{4\gamma_2 C^2 n}$ and $y = 2a_0 r$. The bound is valid whenever $y \geq \gamma_3 C$. If the lower bound is considered as a function of y only, then it is maximized by setting $y = (5L/2)^{2/3}$, and takes the value

$$\frac{1}{a_0}\gamma_1 C \cdot \frac{3}{5}\left(\frac{2}{5L}\right)^{2/3} = \frac{3}{20a_0}\left(\frac{12C}{25n}\right)^{1/3}.$$

The only minor inconvenience is given by the factor a_0, which must be chosen such that the solution r of $2a_0 r = y$ is integer.

Let b_0 be another constant greater than 1, and set $r =$ ceiling of $(y/2b_0)$. Thus,

$$a_0 = \frac{1}{2}\frac{y}{\text{ceiling}(y/2b_0)} \in \left[b_0, b_0/(1 - 2b_0/y)\right] \subseteq \left[b_0, b_0/(1 - u_0)\right]$$

when $y \geq 2b_0/u_0$. Our lower bound now is

$$\frac{3}{20}\frac{1 - u_0}{b_0}\left(\frac{12C}{25n}\right)^{1/3}$$

subject to the conditions

(i) $y \geq \gamma_3 C$ (which is equivalent to $n \geq \dfrac{3C}{50}$);

(ii) $y \geq 2b_0/u_0$.

In addition, the construction of a Lipschitz (C) function f_0 on $[0, \frac{1}{2}(1 - 1/a_0)]$ as in Theorems 5 and 6 will give us a lower bound for C. We can consider a two-piece linear curve with breakpoint at $\frac{1}{4}(1 - 1/a_0)$ (where it takes the value $\frac{3}{2}$) and endpoints at 0 and $\frac{1}{2}(1 - 1/a_0)$ (where the values are 0 and 1 of course). We verify easily that

$$\int_0^{(1/2)(1-1/a_0)} f_0 = \frac{1}{2}\left(1 - \frac{1}{a_0}\right)$$

and that f_0 is Lipschitz (C) when $\frac{1}{4}(1 - 1/a_0)C \leq \frac{3}{2}$, that is, $C \geq 6/(1 - 1/a_0)$. The latter condition is always satisfied when

(iii) $C \geq \dfrac{6}{(1 - 1/b_0)}$.

Take $b_0 = \frac{12}{11}$, $u_0 = \frac{1}{22}$. Condition (iii) gives $C \geq 72$. Condition (ii) leads to $n \geq 12 \cdot 48^3/(25C^2)$. For $C \geq 72$, this holds whenever $n \geq 10$. Replacing u_0 and b_0 in the lower bound has the effect of decreasing the coefficient $\frac{3}{20}$ to $\frac{21}{160}$. This concludes the proof of the first part of Theorem 7.

Before we proceed with the proof of the second part, we should mention that we have struck a compromise between an asymptotically small lower bound and a lower bound that is useful for small values of n. Interested readers can now, with very little extra effort, produce their own bounds for their range of interest. For example, the lower bound for C can be reduced by introducing a "flat level" for f_0 that is much larger than 1. This requires changes in all the theorems, including Theorem 5, but may well be worthwhile. The small optimization problem solved in obtaining the bound for $W(0,1,C)$ can be mimicked for $W(1,1,C)$, but we will not do so here. Instead, we will use the straightforward technique in which we determine $a_0 r$ from $n\int_A g^2 = \frac{1}{8}$, and plug these values back into the formula $\frac{1}{2}r\int_A |g|$.

For $W(1,1,C)$ we start with the function g_0 on $[0,1]$ defined by

$$g_0(x) = \begin{cases} Cx^2/2, & 0 \le x \le \frac{1}{4}, \\ \frac{1}{2}\left(\frac{1}{8}C - C\left(x - \frac{1}{2}\right)^2\right), & \frac{1}{4} \le x \le \frac{3}{4}, \\ C(1-x)^2/2, & \frac{3}{4} \le x \le 1. \end{cases}$$

Clearly, $g_0^{(1)}$ is Lipschitz (C), and g_0 fulfills all the requirements imposed on it in the proof of Theorem 6. Next, $\int|g_0| = C/32$, $\int g_0^2 = 23C/(60 \cdot 256)$, $\sup g_0 = C/16$, so that we can take $\gamma_1 = 1/32$, $\gamma_2 = 23/(256 \cdot 60)$, and $\gamma_3 = 1/16$ in Theorem 6.

From this, we compute, as in Theorem 6, the constants

$$c_3 = \frac{1}{32}\left(\frac{30}{23}\right)^{2/5}, \qquad c_2 = \frac{15}{368}.$$

Thus, this part of Theorem 7 follows directly from Theorem 6. For the sake of the user, we will compute the constant c_1 explicitly. We have taken $a_0 = \frac{1}{2}$, and must make a connection between $(0,0)$ and $(\frac{1}{2}(1 - 1/a_0), 1) = (\frac{1}{4}, 1)$ by using a function f_0 within $W(1,1,C)$, and with average value 1 in this interval. We claim that this can be done for $C \ge 288$. We begin with a basic building block of width $2a$ on which we define the function.

$$\begin{cases} \frac{1}{2}Cx^2, & 0 \le x \le a, \\ Ca^2 - \frac{1}{2}C(x - 2a)^2, & a \le x \le 2a. \end{cases}$$

This function has a Lipschitz (C) derivative on $[0, 2a]$, and has zero derivatives at both endpoints. Its area (integral from 0 to $2a$) is Ca^3, and the maximal value, at $2a$, is Ca^2. If we take a similar function, upside down,

and with a constant D instead of C, where $D \le C$, and stick it to the right of the first function, and if for the new piece we choose a certain width b, then we have our function f_0 provided that we can solve the following equations:

(i) $2a + 2b = \frac{1}{4}$;

(ii) $Ca^3 + Db^3 + 2b = 2a + 2b$ (average value condition);

(iii) $Ca^2 - Db^2 = 1$ (endpoint must be correct);

subject to $D \le C, a \ge 0, b \ge 0$. The solution a is given by

$$\frac{4}{C}\left(1 + \sqrt{1 + \frac{C}{16}}\right).$$

This is less than or equal to $\frac{1}{8}$ (and thus, $b \ge 0$) when $C \ge 128$. Finally, $D = 64 \cdot 8a/(1 - 8a)^2 \le C$ if and only if $C/64 \ge \sqrt{1 + C/16}$, and the latter condition is fulfilled for $C \ge 288$. This concludes the proof of Theorem 7.

Proof of Theorem 8. From Theorem 5, we have the lower bound $(r/2)\varepsilon \int |g|$, subject to $n\varepsilon^2 \int g^2 \le \frac{1}{8}$ if we only look at the subfamily of $Q_r(g)$ with fixed ε. Now, take $\varepsilon = (8n \int g^2)^{-1/2} \le 1$, and we obtain the first inequality. The second bound follows from $\int |g| = \int g^2 = 1/r$ for the given choice of g. The third bound follows from the second bound after taking $r = 8n$. Note that the first inequality would also have been obtained if we had used the stronger half of Theorem 5 and carried out an optimization process as in the proof of Theorem 7.

Proof of Theorem 9. We will apply the general form of Theorem 5. The family $f(b, .)$ is obtained from a central function $f_0(x) = B(1 - 2i\delta)$, $(i - 1)/rB \le x \le i/rB$, $i = 1, \ldots, r$, and $f_0(x) = 0$ elsewhere on $[0, 1]$. Here $\delta = 1/2(2r + 1)$. The area under this function, its integral from 0 to $1/B$ thus, is

$$\sum_{i=1}^{r} (B - 2i\delta)(rB)^{-1} = 1 - \frac{\delta(r + 1)}{B} = 1 - \frac{r + 1}{2B(2r + 1)}.$$

Take now a small perturbation, such as $g(x) = \delta B$ on $[0, 1/2rB)$, $g(x) = -\delta B$ on $[1/2rB, 1/rB)$. Its integral from 0 to $1/rB$ is 0. Since it varies from δB to $-\delta B$, we can add it piecewise to f_0 without violating the monotonic-

ity: define $f(b,.)$ on $[0,1/B)$ as follows:

$$f(b,x) = \begin{cases} f_0(x) + g\left(x + \dfrac{i-1}{rB}\right) & \text{on } \left[\dfrac{i-1}{rB}, \dfrac{i}{rB}\right) \quad \text{if } b_i = 1, \\ f_0(x) & \text{on } \left[\dfrac{i-1}{rB}, \dfrac{i}{rB}\right) \quad \text{if } b_i = -1. \end{cases}$$

Thus, for any b, $f(b,.)$ is monotone on $[0,1/B)$ and $f(0) \le B$. Now, we will distribute the leftover probability uniformly over $[1/B,1)$. This uniform piece has height

$$\frac{\delta(r+1)/B}{1-1/B} = \frac{\delta(r+1)}{B-1}.$$

The value of $f(b,.)$ on $[0,1/B)$ is at least equal to $B(1-(2r+1)\delta) = B/2$. For monotonicity, we must require that $2\delta(r+1) \le B(B-1)$, or, that $(r+1)/(2r+1) \le B(B-1)$. This is satisfied for all integer r when $B \ge 2$. The construction is now complete. We need only compute the α and γ of Theorem 5. As α we can take $\int_0^{1/rB} |g| = \delta/r = 1/2r(2r+1)$. Also, we can take $\gamma = \delta^2/2r(1-2r\delta)$ as can be seen from the following inequality, valid for all i and b:

$$\begin{aligned} 1 - \int \sqrt{f(b_{i+},x) f(b_{i-},x)} \\ &= \int_{(i-1)/rB}^{i/rB} \left(f_0 - \sqrt{f_0(f_0+g)}\right) \\ &\le \frac{B(1-2i\delta)}{2rB}\left(1 - \sqrt{1 + \frac{\delta B}{B(1-2i\delta)}}\right. \\ &\qquad \left. + 1 - \sqrt{1 - \frac{\delta B}{B(1-2i\delta)}}\right) \\ &\le \frac{1-2i\delta}{2r}\left(\frac{\delta B}{B(1-2i\delta)}\right)^2 \\ &\quad \left(\text{by the inequality } \sqrt{1+u} + \sqrt{1-u} \ge 2 - u^2, |u| \le 1\right) \\ &= \frac{\delta^2}{2r(1-2i\delta)} \le \frac{\delta^2}{2r(1-2r\delta)}. \end{aligned}$$

The lower bound is

$$\frac{r\alpha}{4}(1-\gamma)^{2n} = \frac{\delta}{4}\left(1 - \frac{\delta^2}{2r(1 - r/(2r+1))}\right)^{2n}$$

$$\geq \frac{\delta}{4}\left(1 - \frac{1}{r}\delta^2\right)^{2n}$$

$$\geq \frac{\delta}{4}\left(1 - \frac{2n\delta^2}{r}\right).$$

If $r = \overline{(n/4)^{1/3}}$, then clearly $r/\delta^2 = 4r(2r+1)^2 \geq 16((n/4)^{1/3})^3 = 4n$, so that the lower bound is greater than or equal to

$$\frac{\delta}{8} = \frac{1}{16(2r+1)} \geq \frac{1}{16\left(2(n/4)^{1/3} + 3\right)},$$

and this concludes the proof of Theorem 9.

REFERENCES

P. Assouad (1983). Deux remarques sur l'estimation, *Comptes Rendus de l'Académie des Sciences de Paris* **296**, pp. 1021–1024.

L. Birgé (1980). Thèse, 3e partie, Université de Paris VII, Paris, France, 1980.

L. Birgé (1983a). Approximation dans les espaces métriques et théorie de l'estimation, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **65**, pp. 181–237.

L. Birgé (1983b). "On estimating a density using Hellinger distance and some other strange facts", Mathematical Sciences Research Institute, Report MSRI 045-83, University of California, Berkeley.

D. W. Boyd and J. M. Steele (1978). Lower bounds for nonparametric density estimation rates, *Annals of Statistics* **6**, pp. 932–934.

J. Bretagnolle and C. Huber (1979). Estimation des densités: risque minimax, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **47**, pp. 119–137.

P. Deheuvels (1977a). Estimation non paramétrique de la densité par histogrammes généralisés, *Revue de Statistique Appliquée* **25**, pp. 5–42.

P. Deheuvels (1977b). Estimation nonparamétrique de la densité par histogrammes généralisés, *Publications de l'Institut de Statistique de l'Université de Paris* **22**, pp. 1–23.

L. Devroye (1983). On arbitrarily slow rates of global convergence in density estimation, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **62**, pp. 475–483.

R. H. Farrell (1967). On the lack of a uniformly consistent sequence of estimators of a density function in certain cases, *Annals of Mathematical Statistics* **38**, pp. 471–474.

R. H. Farrell (1972). On the best obtainable asymptotic rates of convergence in estimation of a density function at a point, *Annals of Mathematical Statistics* **43**, pp. 170–180.

I. A. Ibragimov and R. Z. Khasminskii (1981). *Statistical Estimation: Asymptotic Theory*, Springer-Verlag, New York.

J. Kiefer (1982). "Optimum rates for non-parametric density and regression estimates, under order restrictions," in *Statistics and Probability: Essays in Honor of C. R. Rao*, G. Kallianpur, P. R. Krishnaiah, and J. K. Ghosh Eds., North-Holland, Amsterdam, pp. 419–428.

A. N. Kolmogorov and V. M. Tikhomirov (1961). ε-entropy and ε-capacity of sets in function spaces, *Translations of the American Mathematical Society* **17**, pp. 277–364.

G. M. Maniya (1969). The square error of the density estimate of a multidimensional normal distribution for a given sample, *Theory of Probability and Its Applications* **14**, pp. 149–153.

E. A. Nadaraya (1974). On the integral mean square error of some nonparametric estimates for the density function, *Theory of Probability and Its Applications* **19**, pp. 133–141.

M. Rosenblatt (1971). Curve estimates, *Annals of Mathematical Statistics* **42**, pp. 1815–1842.

A. M. Samarov (1977). Minimax bound on the risk of nonparametric density estimates, *Problems of Information Transmission* **12**, pp. 242–244.

E. V. Slud (1977). Distribution inequalities for the binomial law, *Annals of Probability* **5**, pp. 404–412.

C. J. Stone (1980). Optimal rates of convergence for nonparametric estimators, *Annals of Statistics* **8**, pp. 1348–1360.

C. J. Stone (1983). "Optimal uniform rate of convergence for nonparametric estimators of a density function or its derivatives," Department of Statistics, University of California, Berkeley, preprint.

G. Wahba (1975). Optimal convergence properties of variable knot, kernel and orthogonal series methods for density estimation, *Annals of Statistics* **3**, pp. 15–29.

CHAPTER 5

Rates of Convergence in L_1

1. INTRODUCTION

One expects that with a finite amount of data a given density estimate has built-in limitations, even for the best densities f. To a certain extent, these limitations are captured in the lower bounds of Chapter 4. In this chapter, we want to obtain very precise information about $E(J_n)$ for particular density estimates, such as the kernel and histogram estimates. In particular, we are interested in asymptotic expressions for $E(J_n)$ for these estimates for the case $d = 1$.

The kernel estimate considered here will be written as

$$f_n(x) = (nh)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{1}$$

(Parzen, 1962; Rosenblatt, 1956), where $h = h_n$ is a given sequence of positive numbers, and K is a given density (kernel) satisfying

$$K(x) = K(-x), \quad \text{all } x; \ K \text{ is bounded and has compact support.} \tag{2}$$

This condition will not be repeated. In view of Theorem 3.1, we also need not consider sequences h that violate

$$\lim_{n \to \infty} h = 0; \ \lim_{n \to \infty} nh = \infty. \tag{3}$$

In this chapter we will consider individual rates of convergence for $E(J_n)$. As we will see, these are closely linked to the following quantities:

$$A(K) = \left(\int K^2\right)^{2/5} \left(\int x^2 K\right)^{1/5}$$

and

$$B(f) = \left(\frac{1}{2}\left(\int \sqrt{f}\right)^4 \int |f''|\right)^{1/5}.$$

To simplify the notation, we will use the symbols α and β for $\sqrt{\int K^2}$ and $\int x^2 K$, respectively. Thus, $A(K) = (\alpha^4\beta)^{1/5}$. The quantity $B(f)$ will be used for all densities f belonging to $\mathscr{F}$, the class of functions satisfying the following:

(i) f is absolutely continuous with a.e. derivative f';

(ii) f' is absolutely continuous with a.e. derivative f'';

(iii) f'' is continuous and bounded.

But because we also want information for densities f not in $\mathscr{F}$, the definition of $B(f)$ must be generalized. For *all* f, we define

$$B^*(f) = \left(\frac{1}{2}\left(\int\sqrt{f}\right)^4 \sup_{h>0}\int|(f*\phi_h)''|\right)^{1/5}$$

where $*$ is the convolution operator, ϕ is a density with compact support and four continuous bounded derivatives, $\phi \in \mathscr{F}$, $\phi'' \in \mathscr{F}$, and $\phi_h(x) = (1/h)\phi(x/h)$. We will prove in Lemma 5 that the value $B^*(f)$ is independent of the choice of ϕ, and that for f in $\mathscr{F}$, the two definitions coincide: $B^*(f) = B(f)$.

Finally, to describe the exact asymptotic behavior of $E(J_n)$, it will be necessary to introduce the function

$$\psi(u) = \sqrt{\frac{2}{\pi}}\left(u\int_0^u e^{-x^2/2}\,dx + e^{-u^2/2}\right), \qquad u \geq 0.$$

It is perhaps useful to get a feeling of how ψ varies. We have because of Mills' ratio ($\int_u^\infty e^{-x^2/2}\,dx \leq (1/u)e^{-u^2/2}$), $\psi(u) \geq u$. By inspection, we also obtain $\psi(u) \geq \sqrt{2/\pi}$. Thus,

$$\max\left(u, \sqrt{\frac{2}{\pi}}\right) \leq \psi(u) \leq u + \sqrt{\frac{2}{\pi}};$$

$$\psi'(u) = \sqrt{\frac{2}{\pi}}\int_0^u e^{-x^2/2}\,dx \geq 0 \quad (\psi \text{ is monotone } \uparrow);$$

$$\psi''(u) \geq 0 \quad (\psi \text{ is convex});$$

$$\lim_{u\downarrow 0}\psi(u) = \sqrt{\frac{2}{\pi}}.$$

THEOREM 1. *For all f in $\mathcal{F}$ having compact support, the kernel estimate defined by* (1)–(3) *satisfies*

$$E(J_n) = J(n,h) + o\left(h^2 + (nh)^{-1/2}\right),$$

where

$$J(n,h) = \int \frac{\alpha\sqrt{f}}{\sqrt{nh}}\psi\left(\sqrt{nh^5}\,\frac{\beta|f''|}{2\alpha\sqrt{f}}\right).$$

Also,

$$J(n,h) \le \sqrt{\frac{2}{\pi}}\,\frac{\alpha\int\sqrt{f}}{\sqrt{nh}} + \frac{\beta}{2}h^2\int|f''|.$$

When f has compact support, then

$$E(J_n) \le \sqrt{\frac{2}{\pi}}\,\frac{\alpha\int\sqrt{f}}{\sqrt{nh}} + \frac{\beta}{2}h^2\sup_{a>0}\int|(f*\phi_a)''| + o\left((nh)^{-1/2}\right),$$

where ϕ is as in the definition of $B^*(f)$. In particular,

$$\limsup_{n\to\infty}\inf_{h>0} n^{2/5}E(J_n) \le C^*A(K)B^*(f),$$

where

$$C^* = 5(8\pi)^{-2/5} = 1.3768102\cdots.$$

The upper bound is not exceeded for the following choice of h when f has compact support, and $B^*(f) < \infty$:

$$h = \left[\frac{\alpha}{2\beta}\,\frac{\int\sqrt{f}}{\sup_{a>0}\int|(f*\phi_a)''|}\sqrt{\frac{2}{\pi}}\right]^{2/5} n^{-1/5}.$$

In the upper bound for $J(n,h)$, we detect a bias component (the second term) and a variance component (the first term). Theorem 1 states that for

densities with finite value for $B^*(f)$, $E(J_n)$ decreases at the rate $n^{-2/5}$ if h is chosen in proportion to $n^{-1/5}$. We should note here that under mildly different regularity conditions (not nested with $\mathscr{F}$), Rosenblatt (1979) obtained a similar upper bound with a slightly larger constant, $5/2^{9/5} = 1.435872\cdots$. Although not explicitly stated by him, his argument uses the inequality $\psi(u) \le 1 + u$, which is not best possible.

The proof of Theorem 1, and its various implications are deferred to other sections, so that we may continue here with a description of our basic results. To complement the upper bound of Theorem 1, we offer a lower bound:

THEOREM 2. *For all f, the kernel estimate defined by* (1) *and* (2) *satisfies*

$$\begin{aligned}\liminf_{n\to\infty} \inf_h n^{2/5} E(J_n) &\ge CA(K)B^*(f) \\ &\ge CC_1 A(K) \\ &\ge CC_1C_2 = C_3,\end{aligned}$$

where

$$C = \inf_u \frac{\psi(u)}{u^{1/5}} = 1.028493\cdots \quad \textit{is a universal constant},$$

$$C_1 = \inf_f B^*(f) = \left(\frac{2^9}{3^4}\right)^{1/5} = 1.4459624\cdots,$$

and

$$C_2 = \inf_{\substack{K \\ \text{even density}}} A(K) = \left(\frac{9}{125}\right)^{1/5} = 0.59083538\cdots.$$

Theorem 2 thus gives a universally applicable lower bound, which in view of the upper bound of Theorem 1 cannot be improved upon very much: note that the constant C^* in the upper bound is only about 35% larger than the constant C in the lower bound. In a sense, the lower bound of Theorem 2 gives us more information than what was available from Chapter 4, because it applies to all densities, not just the "worst" densities in certain classes of densities. More importantly, the lower bound can help us decide whether our sample size n is large enough to achieve a certain L_1 error.

The lower bound remains valid when h is a random variable independent of $X_1, \ldots, X_n$. This would be useful when h is automatically estimated from an independent sample, usually of size much smaller than n.

The existence of the universal lower bound $C_3/n^{2/5}$ is a corollary of the observation that $B^*(f)$ is universally bounded from below by C_1. For the proof of this, we refer to Theorem 3. It is worth noting that this infimum over all $B^*(f)$ is attained for the isosceles triangular density. Thus, the isosceles triangular density is the easiest density to estimate with the kernel estimate. This observation is at the basis of the development of the transformed kernel estimate presented in Chapter 9. The interested reader could now directly skip to that chapter without loss of continuity.

The constant $A(K)$ is minimized by Epanechnikov's kernel

$$K(x) = \frac{3}{4}(1 - x^2), \qquad |x| \le 1,$$

and takes the value C_2 (see Bartlett, 1963, and Epanechnikov, 1969; see also Tapia and Thompson, 1978). For a short proof of this fact, see Lemma 18. Strictly speaking, we should call this kernel Bartlett's and not Epanechnikov's, the name commonly used in the general literature. It is known that the values of $A(K)$ for most reasonable even densities K are very close to the minimal value C_2 (see Rosenblatt, 1971; or Deheuvels, 1977).

Because of its importance, we will devote an entire section to the understanding of $B^*(f)$. The key lemmas and proofs are given in another section. The same treatment is then given to the histogram estimate, where, as we will see, the rate is not $n^{-2/5}$ but $n^{-1/3}$.

In Section 9, we consider uniform upper bounds for the expected L_1 error, that is, bounds that can be applied for any value of n. In Section 10, we give upper bounds for $E(J_n)$ of the type obtained in Theorem 1 for densities f that do not have compact support. Finally, in Section 11, we show how for some smooth but usually long-tailed densities, a specially designed kernel estimate can achieve $E(J_n) \le c/\sqrt{n}$. The price paid for this will be steep: the estimate is not convergent for the vast majority of densities, including all densities with compact support.

2. THE FACTOR $B^*(f)$

From Theorems 1 and 2, we conclude that $B^*(f)$ measures the difficulty posed by f when an ordinary kernel density estimate is used. In this section, we will obtain various properties of $B^*(f)$ and its components. Roughly speaking, we can say that $B^*(f)$ has a component ($\int\sqrt{f}$) that measures how heavy the tail of f is, and a component ($\sup_{h>0}\int|(f * \phi_h)''|$) that measures how oscillatory f is. We will look at these components separately, starting with $\int\sqrt{f}$.

The statement that $\int\sqrt{f}$ is small when f has small tails and vice versa requires some explanation. We can start with a generalization of an inequality due to Carlson (1934) (see Beckenbach and Bellman, 1965, pp. 175):

LEMMA 1. *For* any *random variable X with density f on R,*

(i) $\int\sqrt{f} \le \sqrt{2\pi}(\mathrm{Var}(X))^{1/4}$,

(ii) $\int\sqrt{f} \le \inf_{a\in R} C_\varepsilon(E(|X-a|^{1+\varepsilon}))^{1/2(1+\varepsilon)}$,

where $\varepsilon > 0$ is an arbitrary constant and $C_\varepsilon = (8\pi \sin^{-1}(\pi/(1+\varepsilon))\varepsilon^{-\varepsilon/(1+\varepsilon)})^{1/2}$ depends upon ε only.

Proof. Carlson's inequality (1934) for nonnegative functions g on $[0,\infty)$ states that

$$\int_0^\infty g \le \sqrt{\pi}\left(\int_0^\infty g^2\right)^{1/4}\left(\int_0^\infty x^2 g^2\right)^{1/4}.$$

Now, take $g = \sqrt{f}$, let $p = \int_0^\infty f$, $a = \int_0^\infty x^2 f / \int_{-\infty}^\infty x^2 f$. It is clear that this gives

$$\int\sqrt{f} \le \sqrt{\pi}\left(\int x^2 f\right)^{1/4}\left(p^{1/4}a^{1/4} + (1-p)^{1/4}(1-a)^{1/4}\right).$$

The last factor is maximal, for fixed $a \in (0,1)$, when $p = b/(1+b)$ where $b = (a/(1-a))^{1/3}$, and its maximal value for this p is $(a^{1/3} + (1-a)^{1/3})^{3/4}$, and this, in turn, is never greater than $\sqrt{2}$. Thus, we have

$$\int\sqrt{f} \le \sqrt{2\pi}\left(E(X^2)\right)^{1/4}.$$

But since we can always translate X by a constant value, say, $E(X)$, (i) follows directly.

The second inequality will be obtained independently. For $\varepsilon = 1$, it is only twice as large as (i). For constants $b, c > 0$ we always have

$$\begin{aligned}\int\sqrt{f} &\le c\int(1+b|x|^{1+\varepsilon})f + \int_{c(1+b|x|^{1+\varepsilon})f \le \sqrt{f}}\sqrt{f} \\ &\le c + bcE\left(|X|^{1+\varepsilon}\right) + 2\int_0^\infty \frac{1}{c(1+b|x|^{1+\varepsilon})}\,dx \\ &= c + bcE\left(|X|^{1+\varepsilon}\right) + \frac{1}{c}\cdot\frac{1}{b^{1/(1+\varepsilon)}}\cdot\frac{2\pi/(1+\varepsilon)}{\sin(\pi/(1+\varepsilon))}.\end{aligned}$$

When A, B are nonnegative constants, then the function $cA + B/c$ is minimal for $c = \sqrt{B/A}$, and takes the minimal value $2\sqrt{AB}$. This gives the

bound

$$2\left(\left(1 + bE\left(|X|^{1+\varepsilon}\right)\right)b^{-1/(1+\varepsilon)}\frac{2\pi}{1+\varepsilon} \cdot \frac{1}{\sin(\pi/(1+\varepsilon))}\right)^{1/2}.$$

As a function of b, one can easily verify that this is minimal when $b = (\varepsilon E(|X|^{1+\varepsilon}))^{-1}$, and that the minimal value is the upper bound in (ii) for $a = 0$. The infimum over all $a \in R$ is added to the bound because X can always be translated at will. This concludes the proof of Lemma 1.

Note that we come very close to showing that $E(|X|) < \infty$ implies $\int\sqrt{f} < \infty$. That this is not always true is easily seen from the following example: let f be monotone $\downarrow$ on $[0, \infty)$ such that $f(x) \sim (x \log x)^{-2}$ as $x \to \infty$. Clearly, $\int\sqrt{f} = \infty$ but $E(|X|) < \infty$.

On the other hand, we may have $E(|X|) = \infty$, $\int\sqrt{f} < \infty$: for example, let f be the indicator function of a set A. Clearly, $\int\sqrt{f} = 1$. But if we choose A as $U_i[x_i, x_i + 2^{-i}]$, where $\Sigma x_i 2^{-i} = \infty$, then $E(|X|) = \infty$. Thus, small tails guarantee small values for $\int\sqrt{f}$, while the other implication is not true, except under special conditions. For example, we have inverse inequalities such as

$$E(|X|) \le \sup(|x|\sqrt{f}) \int\sqrt{f}.$$

Densities with a regularly varying tail of order r, that is, $\lim_{x\to\infty} f(tx)/f(x) = t^r$, for all $t > 0$, and similarly for the limit as $x \to -\infty$, have finite values for $\int\sqrt{f}$ when $r < -2$. In particular, all densities with an exponentially decreasing tail or tails have $\int\sqrt{f} < \infty$, but the Cauchy density has $\int\sqrt{f} = \infty$. When $-2 < r$, we must $\int\sqrt{f} = \infty$.

Let us now consider the oscillation factor. We will begin by proving that the oscillation factor is nothing else but $\int|f''|$ when $f \in \mathscr{F}$.

LEMMA 2. *For all $f \in \mathscr{F}$, $B^*(f) = B(f)$.*

Proof. From Theorem 2.1, we recall that

$$\lim_{h\downarrow 0} \int|(f * \phi_h)'' - f''| = \lim_{h\downarrow 0} \int|f'' * \phi_h - f''| = 0$$

when $\int|f''| < \infty$. For such f, $\lim_{h\downarrow 0}\int|(f * \phi_h)''| = \int|f''|$. When $\int|f''| = \infty$, we invoke Fatou's Lemma:

$$\liminf_{h\downarrow 0} \int|(f * \phi_h)''| \ge \int \liminf_{h\downarrow 0}|f'' * \phi_h| = \int|f''|$$

because f'' is continuous. Thus, $B^*(f) \ge B(f)$ for all f in $\mathscr{F}$.

Furthermore,

$$\int |(f * \phi_h)''| = \int |f'' * \phi_h| \le \int |f''|,$$

so that $B^*(f) \le B(f)$. This concludes the proof of the Lemma.

We will now prepare ourselves for the proof of the fact that the value of $B^*(f)$ does not depend upon the choice of ϕ. Our energy will not be wasted, because some of the lemmas obtained in the course of our preparation will be very useful later. Lemmas 3 and 4 are partially overlapping with Lemma 22, but for the time being, they are sufficient for our purposes.

LEMMA 3. *For any density K satisfying* (2) *and any $f \in \mathscr{F}$, we have*

$$f * K_h - f = h^2 \tilde{K}_h * f''$$

for some nonnegative function $\tilde{K}$, where $\tilde{K}$ is symmetric, has compact support, and integrates to $\beta/2$:

$$\int \tilde{K} = \frac{1}{2} \int x^2 K = \frac{\beta}{2}.$$

Proof. Consider a fixed point x and a Taylor series expansion about x:

$$f(y) = f(x) + (y - x) f'(x) + \int_x^y (y - z) f''(z)\, dz.$$

Because K is symmetric, we have

$$\begin{aligned} f * K_h - f &= \int \frac{1}{h} K\left(\frac{x - y}{h}\right) \int_x^y (y - z) f''(z)\, dz\, dy \\ &= \int \frac{1}{h} K\left(\frac{x - y}{h}\right) \int \left[I_{y \ge x} I_{x \le z \le y} (y - z) f''(z)\, dz \right. \\ &\qquad \left. + I_{y \le x} I_{x \ge z \ge y} (-(y - z)) f''(z)\, dz \right] dy \\ &= \int \frac{1}{h} K\left(\frac{x - y}{h}\right) \int \left[I_{y \ge x} I_{x \le z \le y} + I_{y \le x} I_{x \ge z \ge y} \right] \\ &\qquad \times |y - z| f''(z)\, dz\, dy \\ &= h^2 \int \frac{1}{h} \tilde{K}\left(\frac{x - z}{h}\right) f''(z)\, dz, \end{aligned}$$

where

$$h^2 \cdot \frac{1}{h}\tilde{K}\left(\frac{x-z}{h}\right)$$

$$= \int \frac{1}{h}K\left(\frac{x-y}{h}\right)\left[I_{y \geq x}I_{x \leq z \leq y} + I_{y \leq x}I_{x \geq z \geq y}\right]|y - z|\,dy$$

$$= \int K(u)\left[I_{x-hu \geq x}I_{x \leq z \leq x-hu} + I_{x-hu \leq x}I_{x \geq z \geq x-hu}\right]|x - hu - z|\,du$$

$$= h^2\left(\frac{1}{h}\int K(u)\left[I_{u \leq 0}I_{0 \leq (z-x)/h \leq -u} + I_{u \geq 0}I_{0 \geq (z-x)/h \geq -u}\right]\right.$$

$$\left. \times \left|u + \frac{z-x}{h}\right| du\right).$$

It is easy to verify that $\tilde{K}$ is bounded, symmetric, and of compact support when K is. Also,

$$\int \tilde{K} = \int K(u)\left[I_{u \leq 0}I_{0 \leq -x \leq -u} + I_{u \geq 0}I_{0 > -x \geq -u}\right]|u - x|\,du\,dx$$

$$= \int K(u)\left(I_{u \leq 0}\int_u^0 |u - x|\,dx + I_{u \geq 0}\int_0^u |u - x|\,dx\right) du$$

$$= \frac{1}{2}\int u^2 K(u)\,du$$

$$= \frac{\beta}{2},$$

which was to be shown.

LEMMA 4. *Let K be a density satisfying* (2), *and let $\phi \in \mathscr{F}$ be a density with four continuous derivatives and compact support and let $\phi'' \in \mathscr{F}$. Then we have the following inequalities*:

(i) $\int |f * K_h - f| \leq h^2(\beta/2)\int |f''|$, all $f \in \mathscr{F}$, all $h \geq 0$.

(ii) $\int |f * K_h - f| \leq h^2(\beta/2)\liminf_{a \downarrow 0} \int |(f * \phi_a)''|$, all f, all $h \geq 0$.

(iii) $\int |f * K_h - f| \geq h^2\{(\beta/2)\int |(f * \phi_a)''| - h^2\beta_1 \int |(f * \phi_a)''''|\}$, all f, all $h, a > 0$. (Here β_1 is a nonnegative constant depending upon K only.)

Proof. By Lemma 3, the integral in (i) is equal to

$$h^2 \int |\tilde{K}_h * f''| \le h^2 \int \tilde{K} \int |f''| = h^2 \frac{\beta}{2} \int |f''|,$$

where we applied Theorem 2.1. This shows that (i) is true.

Inequality (ii) follows directly from (i) and Theorem 2.1: Observe that

$$\begin{aligned} \int |f * K_h - f| &\le \int |f * K_h - f * \phi_a * K_h| + \int |f - f * \phi_a| \\ &\quad + \int |f * \phi_a * K_h - f * \phi_a| \\ &\le 2 \int |f - f * \phi_a| + h^2 \frac{\beta}{2} \int |(f * \phi_a)''| \end{aligned}$$

and let a tend to 0.

To show (iii), we first consider densities f in $\mathscr{F}$ with four continuous derivatives, $f'' \in \mathscr{F}$, and $\int |f''| < \infty$. By Lemma 3,

$$\begin{aligned} |f * K_h - f| &= |h^2 \Big(\int \tilde{K}_h(x - z) f''(x)\, dz \\ &\quad - \int \tilde{K}_h(x - z)(f''(x) - f''(z))\, dz \Big)| \\ &\ge h^2 \Big(\frac{\beta}{2} |f''(x)| - |\int \tilde{K}_h(x - z)(f''(x) - f''(z))\, dz| \Big). \end{aligned}$$

Reapply Lemma 3 to the last term, but note that f and K should be replaced by f'' and $\tilde{K}$. Thus, there exists a nonnegative constant β_1 depending upon K only such that

$$\int |f * K_h - f| \ge h^2 \Big(\frac{\beta}{2} \int |f''| - h^2 \beta_1 \int |f'''| \Big).$$

For general f, we can apply the inequality just derived to $f * \phi_a$ for all $a > 0$, which together with the fact that

$$\int |f * K_h - f| \ge \int |(f * \phi_a) * K_h - f * \phi_a|$$

yields (iii).

The main result concerning the oscillation factor is captured in the following Lemma:

LEMMA 5. *Let f be an arbitrary density, let K be a density satisfying* (2), *let $\phi \in \mathscr{F}$ be a density with compact support and four continuous derivatives, and let $\phi'' \in \mathscr{F}$. Then the following quantities are equal and independent of K and ϕ:*

$$\lim_{h \downarrow 0} \frac{\int |f * K_h - f|}{h^2 \beta/2} = \liminf_{a \downarrow 0} \int |(f * \phi_a)''| = \sup_{a > 0} \int |(f * \phi_a)''|.$$

Proof. By Lemma 4 (ii), the first factor does not exceed

$$\liminf_{a \downarrow 0} \int |(f * \phi_a)''|.$$

By Lemma 4 (iii), it is at least equal to $\sup_{a>0} \int |(f * \phi_a)''|$. Therefore, the lim inf and the sup are equal, and all three quantities of Lemma 5 are equal to each other.

We have now established a firm connection between the bias of the kernel estimate and our oscillation factor. Let us now show that $B^*(f)$ is bounded from below by a universal constant. To give the reader some insight into the argument used to obtain such lower bounds, we proceed very slowly. First, we will show that $B^*(f) \geq 1$ for all f in $\mathscr{F}$. Then we will show that it is at least 1 for all f, and finally, we will show that the lower bound can be improved to $(2^9/81)^{1/5}$, and that the latter lower bound is attained for the isosceles triangular density.

LEMMA 6. *For all f in $\mathscr{F}$, $B^*(f) \geq 1$.*

Proof. We can assume that $\int |f''| < \infty$. Since f' is absolutely continuous,

$$f'(y) - f'(x) = \int_x^y f''(z)\, dz, \quad \text{all } x < y.$$

Because f'' remains bounded, f' is Lipschitz. Also, since f'' is absolutely integrable, $f'(x)$ tends to 0 as $|x| \to \infty$. Thus, using $(\ )_+$ and $(\ )_-$ for the positive and negative parts of a function, we have

$$\int_{-\infty}^{+\infty} (f''(y))_-\, dy \leq f'(x) \leq \int_{-\infty}^{+\infty} (f''(y))_+\, dy, \quad \text{all } x,$$

and

$$\int_{-\infty}^{+\infty} (f''(y))_+ \, dy + \int_{-\infty}^{+\infty} (f''(y))_- \, dy = 0,$$

so that we may conclude that

$$\sup|f'(x)| \leq \frac{1}{2} \int_{-\infty}^{+\infty} |f''(y)| \, dy.$$

But we also have $1 = \int f \leq \sqrt{\sup f} \int \sqrt{f}$. Combining these inequalities shows that $B(f)^5 \geq \sup|f'(x)|/\sup f^2(x)$. Clearly, by a geometrical argument,

$$1 = \int f(x) \, dx \geq \int (\sup f(x) - |y| \sup|f'(x)|)_+ \, dy$$

$$= \frac{\sup f^2(x)}{\sup|f'(x)|},$$

and the proof of Lemma 6 is complete.

LEMMA 7. *For all f, $B^*(f) \geq 1$.*

Proof. The following facts will be used in this proof:

A. $\int \sqrt{f} = \int \sqrt{f} * \phi_a$ (where ϕ and a are as in the definition of $B^*(f)$).

B. $\int \sqrt{f} * \phi_a \cdot \sup(\sqrt{f} * \phi_a) \geq \int (\sqrt{f} * \phi_a)^2$.

C. $\liminf_{a \downarrow 0} \int (\sqrt{f} * \phi_a)^2 \geq \int \liminf_{a \downarrow 0} (\sqrt{f} * \phi_a)^2$ (Fatou's Lemma)

$= \int (\sqrt{f})^2$ (because $\sqrt{f} * \phi_a \to \sqrt{f}$ at almost all x; see Theorem 2.3)

$= 1$.

D. $f * \phi_a \in \mathscr{F}$.

E. $\sup|(f * \phi_a)'| \leq \frac{1}{2} \int |(f * \phi_a)''|$ (see proof of Lemma 6).

F. $1 \geq \sup(f * \phi_a)^2 / \sup|(f * \phi_a)'|$ (see proof of Lemma 6).

G. $(\sqrt{f} * \phi_a)^2 \leq f * \phi_a$ (Jensen's inequality).

Thus, for fixed a,

$$B^*(f)^5 \geq \left(\int(\sqrt{f} * \phi_a)^2\right)^4 \frac{\sup(f * \phi_a)^2}{\sup(\sqrt{f} * \phi_a)^4} \quad \text{(by A, B, E, and F)}$$

$$\geq \left(\int(\sqrt{f} * \phi_a)^2\right)^4 \quad \text{(by G)}.$$

Take the limit infimum as $a \downarrow 0$, and apply C.

THEOREM 3. *For all f, $B^*(f) \geq (2^9/81)^{1/5}$. The lower bound is attained for the isosceles triangular density.*

Proof. Assume first that $f \in \mathscr{F}$, and that $\int|f''| < \infty$. K and ϕ will be as in (2) and the definition of $B^*(f)$. We have $f'(x) \to 0$ as $|x| \to \infty$. Thus, if f'' is decomposed into its positive and negative parts, $f'' = f''_+ + f''_-$, we have

$$\int|f''| = 2\int f''_+ = -2\int f''_- \geq 2(\sup f' - \inf f').$$

Thus, $B^*(f)^5 \geq (\int\sqrt{f})^4(\sup f' - \inf f')$. We introduce a class of densities containing all densities in $\mathscr{F}$: $\mathscr{G}$ is the class of densities satisfying

$$-D(y - x) \leq f(y) - f(x) \leq C(y - x), \quad \text{all } x < y.$$

Here C and D are positive constants. Among these densities, $\int\sqrt{f}$ is minimized by the triangular density of height b and base split into c and d at the mode, where $b/c = C$, $b/d = D$ and $bc + bd = 2$. Thus, $c^2 = 2/(C + C^2/D)$ and $d^2 = 2/(D + D^2/C)$. For f in $\mathscr{G}$ we have $\sup f' \leq C$ and $\inf f' \geq -D$. Thus, $B^{*5}(f) \geq \inf_{C,D} \inf_{g \in \mathscr{G}}(C + D)(\int\sqrt{g})^4$. Since $(C + D)(\int\sqrt{g})^4$ is scale invariant, we can take $C = 1$. For the triangular density mentioned above, we have

$$\left(\int\sqrt{g}\right)^4(1 + D) \geq \left(\frac{2}{3}\right)^4(c^{3/2} + Dd^{3/2})^4(1 + D)$$

$$= \left(\frac{2}{3}\right)^4(1 + D)\left(\left(\frac{8D^3}{(1 + D)^3}\right)^{1/4} + \left(\frac{8D^2}{(D + D^2)^3}\right)^{1/4}\right)^4$$

$$= \left(\frac{2}{3}\right)^4\frac{8(1 + D)^2}{D} \geq 2\left(\frac{4}{3}\right)^4$$

by noting that $(1 + D)^2/D$ is minimal for $D = 1$. Thus, $\inf_{f \in \mathcal{F}} B^{*5}(f) \geq 2(\frac{4}{3})^4$.

Let f be a density with compact support, and let K be a density in $\mathcal{F}$ with compact support. Let f_n be $f * K_{1/n}$. We will show that $B^*(f_n) \leq B^*(f)(1 + o(1))$, thereby establishing the result for all densities with compact support, since each f_n has compact support and is in $\mathcal{F}$. First, by Fatou's lemma and Theorem 2.3,

$$\liminf_{n \to \infty} \int \sqrt{f * K_{1/n}} \geq \int \sqrt{\liminf_{n \to \infty} f * K_{1/n}} = \int \sqrt{f}.$$

Also, if T is a large enough compact set containing the support of f, we have for n large enough,

$$\int \sqrt{f_n} \leq \int \sqrt{|f_n - f|} + \int \sqrt{f} \leq \sqrt{\lambda(T)} \sqrt{\int |f_n - f|} + \int \sqrt{f} = o(1) + \int \sqrt{f},$$

where we used Theorem 2.1. Next, because for all densities ϕ in $\mathcal{F}$ with compact support, and all $h > 0$,

$$\int |(f * K_{1/n} * \phi_h)''| = \int |(f * \phi_h)'' * K_{1/n}| \leq \int |(f * \phi_h)''|,$$

we see that $B^*(f_n) \leq (1 + o(1))B^*(f)$.

Having established the result for all densities with compact support, we need only standard analytical arguments to generalize it to all densities. This can be done, for example, by approximating f by a sequence of functions g_t, where $g_t = 1$ on $[-t, t]$, 0 outside $[-t - 1, t + 1]$, and smooth and continuous inbetween. (Note that the convolution approach is not applicable since there are densities for which $\int \sqrt{f} < \infty$, yet $\int \sqrt{f * K_h} = \infty$ for all densities K and all $h > 0$.)

3. PROOFS OF THEOREMS 1 AND 2

We start this section by explaining how the function ψ crept into the expression for $J(n, h)$ in Theorem 1. This will be done in two separate important lemmas. The remainder of the proof of Theorem 1 rests on some results about the bias and the variance of the kernel estimate.

LEMMA 8. *Let $X_1, \ldots, X_n$ be independent random variables with a common distribution. Let $E(X_1) = 0$, $E(X_1^2) = \sigma^2 > 0$, and $\rho = E(|X_1|^3) < \infty$. Then,*

$$\sup_{a \in R} |E\left(|(\sigma\sqrt{n})^{-1} \sum_{i=1}^{n} X_i - a| \right) - E(|N - a|)| \leq \frac{c\rho\sigma^{-3}}{\sqrt{n}},$$

where c is a universal positive constant and N is a normal $(0,1)$ random variable. Observe that

$$E(|N - a|) = |a|P(|N| \leq |a|) + \sqrt{\frac{2}{\pi}}\, e^{-a^2/2} = \psi(|a|).$$

Proof. Let F_n be the distribution function of $X = (\sigma\sqrt{n})^{-1}\sum_{i=1}^{n} X_i$, and let Φ be the distribution function of N. Clearly,

$$E(|X - a|) = \int_0^\infty P(|X - a| > t)\, dt = \int_0^\infty (1 - F_n(a + t) + F_n(a - t))\, dt,$$

and a similar equation is valid for N and Φ. The absolute value of the difference between both equations does not exceed

$$\int_0^\infty |\Phi(a + t) - F_n(a + t)|\, dt + \int_0^\infty |\Phi(a - t) - F_n(a - t)|\, dt$$

$$= \int_{-\infty}^{\infty} |\Phi(t) - F_n(t)|\, dt.$$

By well-known nonuniform estimates in the Berry–Esseen type central limit theorem (see Petrov, 1975, Theorem 14, p. 125),

$$|\Phi(t) - F_n(t)| \leq \frac{c\rho\sigma^{-3}}{(1 + |t|^3)\sqrt{n}}$$

for some universal constant c. Since $(1 + |t|^3)^{-1}$ is integrable, we obtain the desired result.

For the expression of $E(|N-a|)$, we note that for $a > 0$,

$$E(|N-a|) = E(|N|) + E(|N-a| - |N|) = E(|N|) + aP(N < 0)$$
$$+ E\big((a-2N)I_{[0<N<a]}\big) - aP(N > a)$$
$$= E(|N|) + a - 2E\big(NI_{[0<N<a]}\big) - 2aP(N > a)$$
$$= \sqrt{2/\pi} + a - aP(|N| > a) - 2\int_0^a \frac{te^{-t^2/2}}{\sqrt{2\pi}}\,dt$$
$$= \sqrt{2/\pi} + aP(|N| < a) - \frac{2\big(1-e^{-a^2/2}\big)}{\sqrt{2\pi}},$$

which was to be shown.

In the remainder of this section, T is an arbitrary interval, $[-r, r]$ is the support of K, K^* is an upper bound for K, and T^* is defined as $\{x: |x-y| \le hr$ for some $y \in T\}$. Thus T^* depends upon h. Also, c is the constant of Lemma 8, $B_n(x) = E(f_n(x)) - f(x)$ is the *bias* at x, $V_n(x) = f_n(x) - E(f_n(x))$ is the *variation* at x, and $\sigma_n^2(x) = E(V_n^2(x))$ is the *variance* at x.

LEMMA 9.

$$\left| E(|f_n(x) - f(x)|) - \sigma_n(x)\psi\left(\frac{|B_n(x)|}{\sigma_n(x)}\right)\right| \le \frac{cK^*}{nh},$$

for all densities K satisfying (2).

Proof. Apply Lemma 8 to the random variables

$$Y_i = \frac{1}{h}K\left(\frac{X_i - x}{h}\right) - E\left(\frac{1}{h}K\left(\frac{X_i - x}{h}\right)\right),$$

and use $a = B_n(x)/\sigma_n(x)$. We obtain an error term in Lemma 8 of the form

$$c\frac{E\big(|Y_1|^3\big)}{nE\big(Y_1^2\big)} \le \frac{cK^*}{nh}.$$

LEMMA 10. *Let T be a bounded interval. Then, for all $h > 0$ and all*

densities K,

$$-\frac{\sqrt{h}}{\alpha}-\sqrt{\int|f*K_h^\dagger-f|\lambda(T)} \le \frac{\sqrt{nh}}{\alpha}\int_T\sigma_n-\int_T\sqrt{f}$$

$$\le \sqrt{\int|f*K_h^\dagger-f|\lambda(T)}$$

where $K^\dagger = K^2/\int K^2$. *Also,* $\int_T|\sigma_n\sqrt{nh}/\alpha-\sqrt{f}| = o(1)$ *as* $h\downarrow 0$.

Proof. We first note that for bounded sets, we always have $\int_T\sqrt{f} < \infty$. We have $\sigma_n^2(x) = a(x) + b(x)$, where

$$a(x) = \frac{\alpha^2 f(x)}{nh}$$

and

$$b(x) = \frac{(f*K_h^\dagger-f)\alpha^2}{nh} - \frac{(f*K_h)^2}{n}.$$

Clearly, $a(x) \ge 0$. Thus, $\sqrt{a(x)+b(x)} \le \sqrt{a(x)} + \sqrt{b_+(x)}$, and $\sqrt{a(x)+b(x)} \ge \sqrt{a(x)} - \sqrt{|b(x)|}$. Integrating over T and applying the Cauchy—Schwarz inequality gives

$$\int_T\sigma_n \le \frac{\alpha}{\sqrt{nh}}\left(\int_T\sqrt{f} + \int_T\sqrt{|f*K_h^\dagger-f|}\right)$$

$$\le \frac{\alpha}{\sqrt{nh}}\left(\int_T\sqrt{f} + \sqrt{\int|f*K_h^\dagger-f|\lambda(T)}\right),$$

and

$$\int_T\sigma_n \ge \frac{\alpha}{\sqrt{nh}}\left(\int_T\sqrt{f} - \sqrt{\int|f*K_h^\dagger-f|\lambda(T)} - \frac{\sqrt{h}}{\alpha}\int f*K_h\right).$$

The first half of Lemma 10 follows easily from this. The last statement of Lemma 10 follows if $\int_T\sqrt{|b(x)|} = o(1)$. But this is a consequence of $h = o(1)$, $\lambda(T) < \infty$, and $\int|f*K_h^\dagger - f| = o(1)$ (see Theorem 2.1).

LEMMA 11. *For all* $f\in\mathscr{F}$, *all K satisfying* (2), *and all bounded intervals T,*

$$\int_T\left||B_n| - \frac{\beta}{2}h^2|f''|\right| = o(h^2) \quad \text{as } h\downarrow 0.$$

The same remains true if T is replaced by R whenever f has compact support, $f \in \mathscr{F}$, and K satisfies (2).

Proof. Let $\tilde{K}$ be the function of Lemma 3. Then

$$\int_T \left| |B_n| - \frac{\beta}{2} h^2 |f''| \right| = \frac{\beta}{2} h^2 \int_T \left| \left| \frac{2}{\beta} \tilde{K}_h * f'' \right| - |f''| \right|$$

$$\leq \frac{\beta}{2} h^2 \int_T \left| \frac{2}{\beta} \tilde{K}_h * f'' - f'' \right|$$

$$= o(h^2)$$

by the Lebesgue dominated convergence theorem. Here we used the fact that $(2/\beta)\tilde{K}_h * f'' \to f''$ at all x (Theorem 2.3), that $|f''|$ is bounded, and that $|(2/\beta)\tilde{K}_h * f''| \leq \sup_x |f''(x)|$.

We need one last technical Lemma before we can attack Theorem 1.

LEMMA 12. *For nonnegative numbers u, v, w, z, we have*

$$\left| u\psi\left(\frac{v}{u}\right) - w\psi\left(\frac{z}{w}\right) \right| \leq |v - z| + \sqrt{\frac{2}{\pi}} |u - w|.$$

Proof. We verify first that $0 \leq \psi'(u) \leq 1$, all $u \geq 0$, and that for all $v \geq 0$, $|(u\psi(v/u))'| \leq \sqrt{2/\pi}$. Thus,

$$\left| u\psi\left(\frac{v}{u}\right) - w\psi\left(\frac{z}{w}\right) \right| \leq \left| u\psi\left(\frac{v}{u}\right) - u\psi\left(\frac{z}{u}\right) \right| + \left| u\psi\left(\frac{z}{u}\right) - w\psi\left(\frac{z}{w}\right) \right|$$

$$\leq |v - z| + \sqrt{\frac{2}{\pi}} |u - w|.$$

Proof of Theorem 1. Theorem 1 has several components. First, we assume that $f \in \mathscr{F}$ and that f has compact support contained in a bounded interval T. Take T so large that for every x in the support of f, the interval $[x - a, x + a]$ is contained in T, where a is a number sufficiently large so that $K_h(u) = 0$ for all n and all $|u| > a$.

We will begin with the inequality of Lemma 12 applied in the following manner:

$$u = \sigma_n(x); \qquad v = |B_n(x)|;$$

$$w = \frac{\alpha\sqrt{f(x)}}{\sqrt{nh}}; \qquad z = \frac{\beta}{2} h^2 |f''(x)|.$$

Now,

$$\int_T |v - z| = o(h^2) \quad \text{(Lemma 11)}$$

and

$$\int_T |u - w| = o\big((nh)^{-1/2}\big) \quad \text{(Lemma 10)}.$$

Thus, combining this into the inequality of Lemma 9 gives

$$\left|\int_T E(|f_n - f|) - J(n,h)\right|$$

$$\leq \left|\int_T E(|f_n - f|) - \int \sigma_n \psi\left(\frac{|B_n|}{\sigma_n}\right)\right| + \left|\int \sigma_n \psi\left(\frac{|B_n|}{\sigma_n}\right) - J(n,h)\right|$$

$$\leq \frac{cK^*}{nh}\lambda(T) + o(h^2) + o\big((nh)^{-1/2}\big),$$

where we used the fact that $J(n,h) = \int_T w\psi(z/w)$.

The inequality involving $J(n,h)$ follows from $\psi(u) \leq u + \sqrt{2/\pi}$:

$$J(n,h) = \int_T w\psi\left(\frac{z}{w}\right) \leq \int_T z + \sqrt{\frac{2}{\pi}} \int_T w.$$

Let us turn now to all densities f having compact support, and let us denote the quantity $\sup_{a>0} \int |(f * \phi_a)''|$ appearing in the definition of $B^*(f)$ by L. Again, from Lemma 9 and the inequality $\psi(u) \leq u + \sqrt{2/\pi}$, we obtain

$$\int_T E(|f_n - f|) \leq \int_T \left(\sqrt{\frac{2}{\pi}}\,\sigma_n + |B_n|\right) + \frac{cK^*}{nh}\lambda(T),$$

and by Lemmas 4, 5, and 10, this is further bounded from above by

$$\sqrt{\frac{2}{\pi}}\,\frac{\alpha}{\sqrt{nh}} \int \sqrt{f} + \sqrt{\frac{2}{\pi}}\,\frac{\alpha}{\sqrt{nh}} \sqrt{\int |f * K_h^\dagger - f| \lambda(T)} + \frac{\beta}{2} h^2 L + \frac{cK^*}{nh}\lambda(T),$$

where $K^\dagger$ is the density defined in Lemma 10. The second term is $o((nh)^{-1/2})$ when $h = o(1)$ (Theorem 2.1). The last term is $o((nh)^{-1/2})$ when $nh \to \infty$. This proves the first upper bound for general f. If we take

the value of h given in the statement of the theorem (i.e., the value that minimizes the main term in the upper bound), then

$$\sqrt{\frac{2}{\pi}}\,\frac{\alpha}{\sqrt{nh}}\int\sqrt{f}+\frac{\beta}{2}h^2L=C^*A(K)\frac{B^*(f)}{n^{2/5}},$$

and this concludes the proof of Theorem 1.

Proof of Theorem 2. We have

$$\inf_h E(J_n)\geq\min\Big(\inf_{h\sqrt{n}\leq 1}E(J_n),\inf_{h\sqrt{n}\geq 1}E(J_n)\Big). \tag{4}$$

Consider a sequence h such that $E(J_n)\sim\inf_h E(J_n)$. It is clear that $E(J_n)\to 0$ for all f, because (3) is sufficient for $E(J_n)\to 0$ (Theorem 3.1). But because $E(J_n)\geq\int|f*K_h-f|$, we must have $h\to 0$ (Theorem 2.4). We will now treat each infimum in (4) separately.

First, if h is such that $h\geq 1/\sqrt{n}$ for all n, and $E(J_n)\sim\inf_{h\sqrt{n}\geq 1}E(J_n)$, then by what we mentioned above, $h\to 0$. Also, $nh\to\infty$, and, in fact, $nh/n^{2/5}\to\infty$. Now, let T be a bounded interval, and $a>0$ be an arbitrary constant. We have for such h the following lower bound for $E(J_n)$:

$$\int_T E(|f_n-f|)\geq\int_T\sigma_n\psi\left(\frac{\int_T|B_n|}{\int_T\sigma_n}\right)-\frac{cK^*}{nh}\lambda(T)$$

(by Lemma 9, the convexity of ψ and Jensen's inequality)

$$\geq C\left(\int_T\sigma_n\right)^{4/5}\left(\int_T|B_n|\right)^{1/5}-o(n^{-2/5})\quad\text{(definition of } C)$$

$$\geq Cn^{-2/5}\left(\alpha\int_T\sqrt{f}\right)^{4/5}\left(\frac{\beta}{2}\int_T|(f*\phi_a)''|\right)^{1/5}(1+o(1))$$

(Lemmas 4 and 10)

$$\sim n^{-2/5}CA(K)\left[\frac{1}{2}\left(\int_T\sqrt{f}\right)^4\int_T|(f*\phi_a)''|\right]^{1/5}$$

(definition of $A(K)$). (5)

Next, let h be a sequence such that $h\leq 1/\sqrt{n}$ for all n, and $E(J_n)\sim\inf_{h\sqrt{n}\leq 1}E(J_n)$. By Theorem 3.1, we know that $nh\to\infty$. Also, by Lemma 3.6,

$$E(J_n)\geq\frac{1}{2}E\left(\int|f_n-f*K_h|\right).$$

By Fatou's lemma,

$$\liminf_{n\to\infty} n^{2/5}E(J_n) \geq \int \frac{1}{2} \liminf_{n\to\infty} n^{2/5}E(|f_n - f * K_h|),$$

and the right-hand side of this is ∞ when for almost all x with $f(x) > 0$, $\liminf_{n\to\infty} n^{2/5}E(|f_n - f * K_h|) = \infty$. To show this, we will use the Berry–Esseen central limit theorem used in Lemma 8. Let $\sigma_n^2(x) = \mathrm{Var}(K_h(X_1 - x))$ and let M be an arbitrarily large positive number. Let Z be a normal $(0, 1)$ random variable. Then,

$$n^{2/5}E(|f_n - f * K_h|) \geq MP\left(|f_n - f * K_h| \geq \frac{M}{n^{2/5}}\right)$$

$$= MP\left(|f_n - f * K_h|\frac{\sqrt{n}}{\sigma_n} \geq \frac{M\sqrt{n}}{\sigma_n n^{2/5}}\right)$$

$$\geq M\left(P\left(|Z| \geq \frac{Mn^{1/10}}{\sigma_n}\right) - 2c\sigma_n^{-3}n^{-1/2}E(|K_h(X_1 - x)$$

$$- E(K_h(X_1 - x))|^3)\right).$$

By inspection of the proof of Lemma 10 and Theorem 2.4, it is easy to see that $\sigma_n^2(x) \sim \alpha^2 f(x)/h$ for almost all x, as $h \to 0$. Also, by the c_r-inequality and Theorem 2.3,

$$E(|K_h(X_1 - x) - E(K_h(X_1 - x))|^3)$$

$$\leq 4E\left(h^{-3}K^3\left(\frac{X_1 - x}{h}\right)\right) + 4(E(K_h(X_1 - x)))^3$$

$$= 4h^{-2}f * (K^3)_h + 4(f * K_h)^3$$

$$\sim 4h^{-2}f(x)\int K^3 + 4f(x)^3 \sim 4h^{-2}f(x)\int K^3, \quad \text{almost all } x.$$

Because $n^{1/10}/\sigma_n(x) \sim n^{1/10}\sqrt{h}/\alpha\sqrt{f(x)} \leq 1/(n^{3/20}\alpha\sqrt{f(x)}) \to 0$ for al-

most all x with $f(x) > 0$, we have

$$n^{2/5}E(|f_n - f * K_h|) \geq M\left(1 - (2c + o(1))\frac{4\int K^3\sqrt{f(x)}}{\sqrt{nh}\,\alpha^3}\right)$$

$$= M(1 - o(1)), \quad \text{almost all } x \text{ with } f(x) > 0.$$

Since M was arbitrary, we have shown that

$$\liminf_{n\to\infty} \inf_{h_1 n \leq 1} n^{2/5}E(J_n) = \infty$$

and this, together with (5), the definition of $B^*(f)$, and the monotone convergence theorem implies

$$\liminf_{n\to\infty} \inf_h n^{2/5}E(J_n) \geq \sup_{\substack{a>0 \\ \text{bounded } T}} CA(K)\left(\int_T \sqrt{f}\right)^{4/5} \frac{\left(\int_T |(f * \phi_a)''|\right)^{1/5}}{2^{1/5}}$$

$$= CA(K)B^*(f).$$

This concludes the proof of Theorem 2.

4. THE HISTOGRAM ESTIMATE

The treatment for the kernel estimate can be mimicked for the histogram estimate on R. We will only consider the simplest histogram estimate defined by the partitions

$$\mathscr{P}_n = \{[kh, (k+1)h), k \text{ integer}\} \tag{6}$$

where $h = h_n$ is a sequence of positive numbers. The L_2 theory for this estimate was thoroughly developed by Freedman and Diaconis (1981) for densities in $\mathscr{F}_2$, that is, functions satisfying

(i) $f \in L_2$, f is absolutely continuous with a.e. derivative f';

(ii) $0 < \int f'^2 < \infty$.

For example, Scott (1979) and Freedman and Diaconis (1981) proved Theorem 4 below.

THEOREM 4. *When $f \in \mathscr{F}_2$, the histogram estimate defined by* (6) *satisfies*

$$\inf_h E\left(\int (f_n - f)^2\right) \sim \frac{\frac{1}{4}\left(36\int f'^2\right)^{1/3}}{n^{2/3}},$$

and this rate is attained if we take $h \sim (6/(n\int f'^2))^{1/3}$.

Theorem 4 is stated here without proof, merely for later reference. We note that for the smoothest f, $n^{-2/3}$ is the optimal L_2 rate as compared to $n^{-4/5}$ for the kernel estimate. For smooth f, we expect a better L_2 performance with the kernel estimate at least for large n. We will show in this section that the same remains true for $E(J_n)$. Of course, one cannot extend this nesting of performances to all densities f. For example, when f is uniform on $[0, 1]$ and $h = 1$, then $E(J_n) = 0$ for all n, while, regardless of how h is chosen,

$$\liminf_{n\to\infty} n^{2/5} E(J_n) = \infty$$

for the kernel estimate (Theorem 2), because $\sup_{a>0} \int |(f * \phi_a)''| = \infty$ and thus $B^*(f) = \infty$.

In this section, $\mathscr{F}$ will denote the class of functions satisfying (i) and (ii):

(i) f is absolutely continuous with a.e. derivative f';

(ii) f' is bounded and continuous. (Note: $\int |f'| < \infty$.)

THEOREM 5. *For all $f \in \mathscr{F}$, the histogram estimate defined by* (6) *satisfies the following lower bound:*

$$\liminf_{n\to\infty} \inf_h n^{1/3} E(J_n) \geq C_H B_H(f) \geq C_H,$$

where

$$C_H = \inf_u \psi(u)/(2u)^{1/3} = 0.880261\cdots$$

and

$$B_H(f) = \left(\frac{1}{2}\left(\int \sqrt{f}\right)^2 \int |f'|\right)^{1/3}$$

(*which is ≥ 1 for all f in $\mathscr{F}$*).

For the present section, and Section 5, we need another function, r_n, defined as follows:

$$r_n(x) = \frac{x - a}{b - a}, \quad \text{all } x \in A_{nj} = [a, b) = [jh, (j + 1)h).$$

This, on the real line, is a sawtooth function taking values between 0 and 1. We will also need

$$z_n(x) = (1 - 2r_n(x))f'(x),$$

a function that is well-defined for f in $\mathscr{F}$. The function z_n oscillates between $-f'$ and $+f'$.

THEOREM 6 [Exact Asymptotic Behavior of $E(J_n)$]. *Let $f \in \mathscr{F}$ have compact support. For the histogram estimate defined by* (6) *and* (3), *we have*

$$E(J_n) = J(n, h) + o\left(h + \frac{1}{\sqrt{nh}}\right),$$

where

$$J(n, h) = \int \sqrt{\frac{f}{nh}}\, \psi\left(\frac{h}{2}|z_n|\sqrt{\frac{nh}{f}}\right).$$

Also,

$$\limsup_{n\to\infty} \inf_h n^{1/3} E(J_n) \le C_H^* B_H(f),$$

where $C_H^* = (27/4\pi)^{1/3} = 1.290381\cdots$.

Thus, on $\mathscr{F}$, we have squeezed $E(J_n)$ between two close bounds, $C_H B_H(f)/n^{1/3}$ and $C_H^* B_H(f)/n^{1/3}$, where $C_H^*/C_H = 1.46590\cdots$. Thus, $B_H(f)$ measures to some extent the difficulty f poses for the histogram estimate. What is intriguing is that $B_H(f) \neq B^*(f)$. In other words, densities that are easy to estimate with the kernel estimate may be difficult to handle with the histogram estimate and vice versa. For example, the uniform [0, 1] density has $B^*(f) = \infty$, while it can be approximated by a sequence in $\mathscr{F}$ with B_H values tending to 1, the minimum possible value for B_H. The rectangular density is thus the easiest density for the histogram estimate.

A second consequence of Theorems 5 and 6 is that, for smooth densities, the average L_1 error for the histogram estimate must vary at least as $n^{-1/3}$, thus at a rate worse than that for the kernel estimate ($n^{-2/5}$). Whether for a particular n the kernel estimate is better depends to a large extent upon the values of $B_H(f)$ and $B^*(f)$.

Distribution-free lower bounds of the type obtained in Theorem 2 for the kernel estimate do not exist for the histogram estimate. It seems that to

obtain such bounds, one should consider criteria of the type

$$\inf_h \sup_{x_0} E(J_n)$$

where $\mathscr{P}_n$ is replaced by $\mathscr{P}_n(x_0) = \{[kh + x_0, (k+1)h + x_0), k \text{ integer}\}$. This will not be pursued here because of the clear asymptotic inferiority of the histogram estimate already established on a large subclass of $\mathscr{F}$.

5. PROOFS OF THEOREMS 5 AND 6

In this section, we let μ_n be the empirical measure for $X_1, \ldots, X_n$, and let μ be the measure corresponding to f. Let $A_{nj} = [jh, (j+1)h)$,

$$f_n(x) = \frac{\mu_n(A_{nj})}{h}, \qquad x \in A_{nj},$$

and

$$g_n(x) = E(f_n(x)) = \frac{\mu(A_{nj})}{h}, \qquad x \in A_{nj}.$$

Throughout the section, T is a bounded interval of the type $[-t, t]$.

LEMMA 13. *If $\int |g_n - f| \to 0$ for a given $f \in \mathscr{F}$, then $\lim_{n\to\infty} h = 0$.*

Proof. We note that Lemma 13 does not hold for all f, and that it is not a consequence of Theorem 2.6.

Assume first that $\lim_{n\to\infty} h = \infty$. Then, $g_n \to 0$ for all x, and $\liminf_{n\to\infty} \int |g_n - f| \geq \int \liminf_{n\to\infty} |g_n - f| = 1$, which is a contradiction. Assume next that $\lim_{n\to\infty} h = c > 0$. For $h = c$, $\int |g_n - f| = 0$ implies that $f' = 0$ almost everywhere. But since f is absolutely continuous, this would imply that $f = 0$ almost everywhere. Thus, we must conclude that for $h = c > 0$, $\int |g_n - f| > 0$. Let us make the dependence upon h explicit as in g_h. It is clear that we have a contradiction, once again, if we can show that $h \to c$ implies $\int |g_h - g_c| \to 0$. This contradiction would conclude the proof of Lemma 13.

To prove this, we need only show that $g_h \to g_c$ for almost all x and apply Theorem 2.8. Assume, without loss of generality, that $x > 0$. Now, $x \in [jh, (j+1)h) \cap [kc, (k+1)c)$ for some integers j, k. For h close enough to c, we have $j = k$. Let A be the intersection of the two intervals,

and let B be the symmetric difference. We have

$$|g_h - g_c| \le \mu(A)\left|\frac{1}{h} - \frac{1}{c}\right| + \mu(B)\left(\frac{1}{h} + \frac{1}{c}\right) \le \frac{|h-c|}{hc} + o(1) = o(1).$$

This concludes the proof of Lemma 13.

LEMMA 14. *If X is a binomial (n, p) random variable, then*

$$\sup_i P(X = i) \le \frac{c}{\sqrt{np(1-p)}}$$

for some universal constant c.

Proof. We use the fact that $k! = (k+1)^k\sqrt{2\pi(k+1)}\,e^{-(k+1)}\exp(\theta/12(k+1))$ for some $0 < \theta < 1$ (see, e.g., Whittaker and Watson, 1963). Thus, since all terms $P(X = i)$ are less than the kth where $k = \lfloor (n+1)p \rfloor$ (Feller, 1968, p. 151), we have

$$P(X = i) \le \left(1 + \frac{1}{n}\right)^n \frac{e^{1+1/24}\sqrt{n+1}}{\sqrt{2\pi(k+1)(n-k+1)}}$$

$$\le \frac{e^{2+1/24}\sqrt{2n}}{\sqrt{2\pi(n+1)p(n+1)(1-p)}}$$

$$\le \frac{c}{\sqrt{np(1-p)}},$$

where we used the inequality $(1+u) \le e^u$.

LEMMA 15.

$$\inf_{f \in \mathcal{F}} B_H(f) = 1.$$

Proof. We know that $\int\sqrt{f} \ge \int f / \sqrt{\sup f} = 1/\sqrt{\sup f}$. Since f is absolutely continuous, $\int |f'| \ge 2 \sup f$. Combining these inequalities shows that

$$\frac{1}{2}\left(\int\sqrt{f}\right)^2\left(\int|f'|\right) \ge 1.$$

To show that this lower bound can be attained, we construct a sequence of densities $f = g * \phi_h$ in $\mathcal{F}$ and let $h \to 0$: here ϕ is the normal $(0, 1)$ density and g is the uniform $[0, 1]$ density.

It is a simple exercise in analysis to show that $\int\sqrt{g * \phi_h} \to \int\sqrt{g} = 1$ as $h \to 0$.

Furthermore, since the convolution of two symmetric unimodal distributions is unimodal (Feller, 1971, pp. 167–168), each f is unimodal. Therefore,

$$\int|(g * \phi_h)'| = 2\sup(g * \phi_h) \to 2\sup g = 2 \quad \text{as } h \to 0.$$

This shows that for this sequence $\limsup B_H(f) \le 1$.

We inherit the notation of Sections 2 and 3. In particular, $\sigma_n^2(x) = \mathrm{Var}(f_n(x)) = \mathrm{Var}(\mu_n(A_{nj})/h)$, $x \in A_{nj}$; and $B_n(x) = g_n(x) - f(x)$. Also,

$$E(J_n) = \sum_j E\big(J_n(A_{nj})\big), \quad \text{where } J_n(A) = \int_A |f_n - f|.$$

In analogy with Lemma 9,

$$\left| E\big(J_n(A_{nj})\big) - \int_{A_{nj}} \sigma_n(x)\psi\left(\frac{|B_n(x)|}{\sigma_n(x)}\right)dx \right| \le \frac{c}{n}, \quad \text{for all } j, \tag{7}$$

where the error term is computed as follows from Lemma 8:

$$c\int_{A_{nj}} \frac{E\big(|I_{A_{nj}}(X_1) - \mu(A_{nj})|^3/h^3\big)}{nE\big(|I_{A_{nj}}(X_1) - \mu(A_{nj})|^2/h^2\big)}\,dx \le \frac{c}{nh}\int_{A_{nj}} dx = \frac{c}{n}.$$

Also, in analogy with (5), we have

$$E(J_n(T)) \ge \int_T \sigma_n \psi\left(\frac{\int_T |B_n|}{\int_T \sigma_n}\right) - \frac{c}{n}\left(\frac{\lambda(T)}{h} + 2\right)$$

$$\ge C_H\left(\int_T \sigma_n\right)^{2/3}\left(2\int_T |B_n|\right)^{1/3} - \frac{c}{n}\left(\frac{\lambda(T)}{h} + 2\right) \tag{8}$$

because $C_H = \inf \psi(u)/(2u)^{1/3}$.

LEMMA 16. *Assume that* $\lim_{n\to\infty} h = 0$. *Then* $\int_T |\sigma_n\sqrt{nh} - \sqrt{f}| = o(1)$.

Proof. Let $x \in A_{nj}$. Then $\sigma_n^2(x) = \mu(A_{nj})(1 - \mu(A_{nj}))/(nh^2)$, and thus

$$\begin{aligned}|\sigma_n\sqrt{nh} - \sqrt{f}| &= |\sqrt{\mu(A_{nj})(1-\mu(A_{nj}))/h} - \sqrt{f}| \\ &\le \frac{\mu(A_{nj})}{\sqrt{h}} + |\sqrt{\mu(A_{nj})/h} - \sqrt{f}| \\ &= g_n\sqrt{h} + |\sqrt{g_n} - \sqrt{f}| \le g_n\sqrt{h} + \sqrt{|g_n - f|}.\end{aligned}$$

Now, take integrals on left and right, and apply the Cauchy–Schwarz inequality and Theorem 2.6:

$$\int_T |\sigma_n\sqrt{nh} - \sqrt{f}| \le \sqrt{h} + \int_T \sqrt{|g_n - f|} \le \sqrt{h} + \sqrt{\int_T |g_n - f| \lambda(T)} = o(1).$$

LEMMA 17. *Assume that* $\lim_{n\to\infty} h = 0$, *and that* $f \in \mathscr{F}$ *has compact support. Then*

$$q_n(x) = o(h), \quad \textit{for all } x, \quad \textit{and} \quad \int q_n = o(h),$$

where $q_n(x) = |B_n(x) - (h/2)z_n(x)|$. *For all* $f \in \mathscr{F}$, *we have*

$$\int_T q_n = o(h) \quad \textit{and} \quad \int_T |B_n| \sim \frac{h}{2}\int_T |z_n| \sim \frac{h}{4}\int_T |f'|.$$

Proof. By Taylor's expansion with remainder,

$$f(y) = f(x) + (y - x)f'(x) + (y - x)(f'(\xi) - f'(x)), \qquad x \le \xi \le y.$$

Thus, for $x \in A_{nj}$,

$$\begin{aligned}g_n(x) &= \frac{1}{h}\int_{A_{nj}} f = f(x) + \frac{1}{h}\int_{A_{nj}} (y - x)f'(x)\,dy \\ &\quad + \frac{1}{h}\int_{A_{nj}} (y - x)(f'(\xi) - f'(x))\,dy \\ &= f(x) + \frac{h}{2}z_n(x) + b(x, h),\end{aligned}$$

where $|b(x, h)| \le \sup_{|y-x|\le h} |f'(y) - f'(x)|h/2$ is bounded by h times an integrable function (because f has compact support, and $\sup|f'| < \infty$), and $b(x, h)/h \to 0$ for all x (because f' is continuous). Thus, $\int q_n \le \int |b(x, h)| = o(h)$, and $q_n(x) = o(h)$ for all x. When f does not have compact support, it remains true that $\int_T q_n = o(h)$.

Thus, for all $f \in \mathscr{F}$, $\int_T |B_n| \sim (h/2)\int_T |z_n|$. We are done if we can show that for such f, $\int_T |z_n| \sim \frac{1}{2}\int_T |f'|$. Now, let N_n be the collection of indices j

for which $A_{nj} \subseteq T$. Obviously

$$\sum_{j \in N_n} \int_{T \cap A_{nj}} |f'||1 - 2r_n| \leq \sup|f'| \cdot 2h = o(1).$$

Because f' is continuous on T, it is uniformly continuous on T. In particular, for fixed $\varepsilon > 0$, we have $|f'(x) - f'(y)| < \varepsilon$ for all $|x - y| \leq h$, $x, y \in T$, and all h small enough. Using this, we have the following chain of inequalities:

$$\begin{aligned}
\sum_{j \in N_n} \int_{T \cap A_{nj}} |f'||1 - 2r_n| &\leq \sum_{j \in N_n} \sup_{T \cap A_{nj}} |f'| \int_{A_{nj}} |1 - 2r_n| \\
&= \sum_{j \in N_n} \sup_{T \cap A_{nj}} |f'| \int_0^h \left|1 - 2\frac{x}{h}\right| dx \\
&= \sum_{j \in N_n} \sup_{T \cap A_{nj}} |f'| \frac{h}{2} \\
&\leq \sum_{j \in N_n} \frac{1}{2} \int_{T \cap A_{nj}} (|f'| + \varepsilon) + O(h) \\
&= \frac{1}{2} \int_T |f'| + \frac{1}{2} \int_T \varepsilon + O(h).
\end{aligned}$$

Similarly, a lower bound $\frac{1}{2}\int_T |f'| - \frac{1}{2}\int_T \varepsilon - O(h)$ is obtained, and we are done.

Proof of Theorem 5. We will split the proof in half by the following device:

$$\inf_h E(J_n) \geq \min\Big(\inf_{h\sqrt{n} \geq 1} E(J_n), \inf_{h\sqrt{n} \leq 1} E(J_n) \Big).$$

First, let h be a sequence such that $E(J_n) \sim \inf_{h\sqrt{n} \geq 1} E(J_n)$. We know that $h \to 0$ (because $E(J_n) \geq \int |g_n - f|$, and this tends to 0 for $f \in \mathscr{F}$ if and only if $h \to 0$, by Theorem 2.6 and Lemma 13). Also, $1/nh = o(n^{-1/3})$.

Thus, combining (8) with Lemmas 16 and 17 gives for all $T = [-t, t]$:

$$\inf_{h\sqrt{n} \geq 1} E(J_n) - E(J_n) \geq E(J_n(T))$$

$$\geq \int_T \sigma_n \psi\left(\frac{\int_T |B_n|}{\int_T \sigma_n}\right) - \frac{c}{n}\left(\frac{\lambda(T)}{h} + 2\right)$$

$$\geq C_H\left(\int_T \sqrt{f}\right)^{2/3}\left(\int_T |f'|\right)^{1/3}(2n)^{-1/3}(1 + o(1)).$$

Now, let T tend to R, and conclude that

$$\liminf_{n\to\infty} \inf_{h\sqrt{n} \leq 1} E(J_n)n^{1/3} \geq C_H B_H(f). \tag{9}$$

Next, we consider a sequence for which $n^{1/3}E(J_n) \sim \inf_{h\sqrt{n} \leq 1} n^{1/3}E(J_n)$. By Fatou's lemma,

$$n^{1/3}E(J_n) \geq \frac{1}{2}\int \liminf_{n\to\infty} n^{1/3}E(|f_n - g_n|).$$

Fix $x \in A_{nj}$, and let Z be a binomial $(n, \mu(A_{nj}))$ random variable. By Lemma 14,

$$n^{1/3}E(|f_n - g_n|) = n^{1/3}(nh)^{-1}E(|Z - E(Z)|)$$

$$\geq MP(|Z - E(Z)| \geq Mhn/n^{1/3}) \quad \text{(for arbitrary } M > 0)$$

$$\geq M\left(1 - \frac{2Mhn}{n^{1/3}}\frac{3}{\sqrt{n\mu(A_{nj})(1 - \mu(A_{nj}))}}\right), \tag{10}$$

where c is the constant of Lemma 14.

By Theorem 2.2, $\mu(A_{nj})/h \to f(x)$ for almost all x because $h \to 0$). Also, $\mu(A_{nj}) \to 0$ for almost all x. Thus, (10) $\sim M(1 - 2Mc\sqrt{hn^{1/3}/f(x)})$ $\sim M$ for almost all x with $f(x) > 0$. Since M was arbitrary, we have

$$\liminf_{n\to\infty} \inf_{h\sqrt{n} \leq 1} n^{1/3}E(J_n) = \infty. \tag{11}$$

Theorem 5 now follows by combining (9) and (11), and applying Lemma 15.

Proof of Theorem 6. We start from inequality (7) and let T be a set $[-t, t]$ containing the support of f. Then

$$\left| E(J_n) - \int \sigma_n \psi(|B_n|/\sigma_n) \right| \leq \frac{c}{n}\left(\frac{\lambda(T) + 2h}{h} + 2\right) = O\left(\frac{1}{nh}\right). \quad (12)$$

Let q_n keep its meaning from Lemma 17 and let p_n be $|\sigma_n - \sqrt{f/nh}|$. In Lemma 16, we have proved that $\int p_n = o(1/\sqrt{nh})$. Arguing as in the proof of Theorem 1,

$$\left|\sigma_n \psi\left(\frac{|B_n|}{\sigma_n}\right) - \sqrt{\frac{f}{nh}}\,\psi\left(\frac{h}{2}|z_n|\sqrt{\frac{nh}{f}}\right)\right| \leq \sigma_n\left|\psi\left(\frac{|B_n|}{\sigma_n}\right) - \psi\left(\frac{h}{2}\frac{|z_n|}{\sigma_n}\right)\right|$$

$$+\left|\sigma_n \psi\left(\frac{h}{2}\frac{|z_n|}{\sigma_n}\right) - \sqrt{\frac{f}{nh}}\,\psi\left(\frac{h}{2}|z_n|\sqrt{\frac{nh}{f}}\right)\right| \leq q_n + \sqrt{\frac{2}{\pi}}\,p_n. \quad (13)$$

Combining (12), (13), Lemma 16, and Lemma 17 shows that $E(J_n) - J(n, h) = o(h + 1/\sqrt{nh})$.

The second part of Theorem 6 uses the inequality $\psi(u) \leq \sqrt{2/\pi} + u$ on $J(n, h)$:

$$J(n, h) \leq \sqrt{\frac{2}{\pi}} \int \sqrt{\frac{f}{nh}} + \frac{h}{2}\int |z_n| \sim \sqrt{\frac{2}{\pi}} \int \sqrt{\frac{f}{nh}} + \frac{h}{4}\int |f'|.$$

Then, choose the h that minimizes this upper bound, that is,

$$h = \left(\frac{8}{\pi}\left(\frac{\int \sqrt{f}}{\int |f'|}\right)^2 \frac{1}{n}\right)^{1/3},$$

and the result follows.

6. CHOICE OF THE SMOOTHING PARAMETER

In Theorems 1 and 6, we obtained exact asymptotic expressions for $E(J_n)$. Unfortunately, the sequence h that minimizes the main term in the asymptotic expressions is hard to extract in closed form. The upper bounds of Theorems 1 and 6 lend themselves better to such minimization. We obtain for the kernel estimate, after choosing the Epanechnikov kernel $\frac{3}{4}(1 - x^2)$, $|x| \leq 1$ (with $\int K^2 = \frac{3}{5}$ and $\int x^2 K = \frac{1}{5}$):

$$h = \left[\sqrt{\frac{15}{2\pi}} \frac{\int \sqrt{f}}{\int |f''|} \right]^{2/5} n^{-1/5} \tag{14}$$

valid for $f \in \mathscr{F}_K$, the class of densities for which Theorem 1 is valid. When $f \in \mathscr{F}_H$, the class of densities for which Theorem 6 holds, the corresponding h for the histogram estimate is

$$h = \left[\sqrt{\frac{8}{\pi}} \frac{\int \sqrt{f}}{\int |f'|} \right]^{2/3} n^{-1/3}. \tag{15}$$

We should stress that (14) and (15) are only valid for $f \in \mathscr{F}_K$ (or $\in \mathscr{F}_H$), and then only for those f with finite $\int \sqrt{f}$ and $\int |f''|$ (or $\int |f'|$). If one of these integrals is infinite, we obtain the strange result that the nearly optimal h is ∞ or 0. This contradiction is of course due to the invalidity of the formal minimizations when one of the terms involved is infinite or 0 (see, e.g., the upper bound for $J(n, h)$ in the proof of Theorem 6). Since all these conditions are hard to check, any attempt at determining h by approximating (14) or (15) is doomed to cause serious concern among the users.

For the choice of h, we also refer to Sections 5.8, 5.9, and 5.10, and Chapters 6 and 9. In Chapters 6 and 9, for example, very general consistency theorems are given for density estimates with data-dependent smoothing factors. Here, we would just like to point out the features of the *parametric method* for determining h.

Assume that f belongs to, or is close to, a member of a parametric family $f_\theta, \theta \in R^k$. For this family, (14) and (15) are explicitly known (by assumption): $c_K(\theta)/n^{1/5}$ and $c_H(\theta)/n^{1/3}$. Estimate θ from the data in a conventional way by $\hat{\theta}$, and use $c_K(\hat{\theta})/n^{1/5}$ and $c_H(\hat{\theta})/n^{1/3}$ instead of (14) and (15). If at all possible, robust estimates should be used for $\hat{\theta}$. This, the

parametric method, was suggested for L_2 optimal values of h for the kernel estimate by Deheuvels (1977) and Deheuvels and Hominal (1980). A similar development for the histogram estimate can be found in Scott (1979) and Freedman and Diaconis (1981). In particular, for the densities satisfying Rosenblatt's condition (Section 4.3), the L_2-optimal h for the kernel estimate is given by

$$h = \left(\frac{15}{n \int f''^2} \right)^{1/5} \tag{16}$$

(valid for the Epanechnikov kernel), and for the densities of Theorem 4, the L_2-optimal h for the histogram estimate is

$$h = \left(\frac{6}{n \int f'^2} \right)^{1/3}. \tag{17}$$

Except for their dependence upon n, there is very little resemblance between (16), (17) and (14), (15). It is curious that even for simple densities such as the normal (0, 1) density, the values are very different. For example, in that case, (16) and (17) give

$$h = \left(\frac{40\sqrt{\pi}}{n} \right)^{1/5} = 2.345 \cdots n^{-1/5}$$

and

$$h = \left(\frac{12\sqrt{\pi}}{n} \right)^{1/3} = 2.7706 \cdots n^{-1/3},$$

respectively. But (14) and (15) correspond to

$$h = \left(\frac{15e\sqrt{2\pi}}{8n} \right)^{1/5} = 1.6644 \cdots n^{-1/5}$$

and

$$h = \left(4\frac{\sqrt{8\pi}}{n} \right)^{1/3} = 2.7168 \cdots n^{-1/3},$$

respectively. (It is easy to verify that (14), (15) remain valid for the normal density.) Of course, more extreme cases can be constructed. For example, when we approach the Cauchy density within the family of Student's t densities, the coefficient of n in (16) tends to a constant, while the coefficient of n in (17) tends to ∞, due to the fact that the L_1 theory is more sensitive to the weight in the tails of the distributions. In Table 1, we give $\int\sqrt{f}$, $\int|f'|$, $\int|f''|$, $B^*(f) = [\frac{1}{2}(\int\sqrt{f})^4\int|f''|]^{1/5}$, $B_H(f) = [\frac{1}{2}(\int\sqrt{f})^2\int|f'|]^{1/3}$, $c_K(\theta)$, and $c_H(\theta)$ for various families of distributions. When $f \notin \mathscr{F}_H$, $\int|f'|$ is replaced by a limit of values of $\int|f'|$ of densities in $\mathscr{F}_H$ tending to f. When $f \notin \mathscr{F}_K$, $\int|f''|$ is replaced by its generalization $\sup_{a>0}\int|(f * \phi_a)''|$.

It is not necessary to mention location and scale parameters in the families (and thus in the expressions for $c_K(\theta)$ and $c_H(\theta)$) because

$$c_K(\mu, \sigma, \gamma) = \sigma c_K(0, 1, \gamma)$$

and

$$c_H(\mu, \sigma, \gamma) = \sigma c_H(0, 1, \gamma),$$

where μ is a location parameter, σ is a scale parameter, and γ is a collection of zero or more shape parameters. The computations of Table 1 are made easier because for unimodal f in $\mathscr{F}_H$, $\int|f'| = 2\sup f$, and for symmetric $f \in \mathscr{F}_K$ with unimodal f' on $[0, \infty)$, $\int|f''| = 4\sup|f'|$.

None of the densities in Table 1 have a shape parameter. Since the scale parameters can be factored out, Table 1 gives us complete information about $c_K(\theta)$ and $c_H(\theta)$. For example, if f is nearly normal, we could take for the histogram estimate

$$h = (\sqrt{8\pi})^{1/3}\hat{\sigma}n^{1/3}$$

where $\hat{\sigma}$ is a data-based robust estimate of σ, under the assumption that the data come from a normal (μ, σ^2) density. In general, such robust estimates can be obtained as follows: choose two numbers $p, q \in (0, 1)$, and let x_1 and x_2 be the (known) quantiles of the distribution with $\mu = 0$, $\sigma^2 = 1$, corresponding to p and q, respectively. Thus, the order statistics $X_{(np)}$ and $X_{(nq)}$ obtained from $X_1, \ldots, X_n$ can be considered estimates of $\mu + \sigma x_1$ and $\mu + \sigma x_2$, respectively. Therefore, σ can be estimated by $(X_{(nq)} - X_{(np)})/(x_2 - x_1)$. Estimates of this kind are usually preferable over estimates that are based on averaging. For example, if we add a scale factor to the Cauchy density of Table 1, and take $p = \frac{1}{4}$, $q = \frac{3}{4}$, we obtain the following estimate for σ: $\frac{1}{2}(X_{(3n/4)} - X_{(n/4)})$ (in the Cauchy case, averaging would have been

TABLE 1

Density	$\int\sqrt{f}$	$\int\lvert f'\rvert$	$\int\lvert f''\rvert$	$B^*(f)$	$B_H(f)$	$c_K(\theta)$	$c_H(\theta)$
Uniform on [0, 1]	1	2^a	∞^b	∞	1 (minimum)	0	$(2/\pi)^{1/3}$
Isosceles triangle on [0, 1]	$\frac{1}{2}(2^{3/2})$	4^a	16^b	$(2^9/3^4)^{1/5}$ (minimum)	$(16/9)^{1/3}$	$(5/192\pi)^{1/5}$	$(4/9\pi)^{1/3}$
Normal (0, 1)	$(8\pi)^{1/4}$	$\sqrt{2/\pi}$	$\sqrt{8/\pi e}$	$(128\pi/e)^{1/10}$	$2^{1/3}$	$(225\pi e^2/32)^{1/10}$	$(4\sqrt{8\pi})^{1/3}$
Laplace $(e^{-\lvert x\rvert}/2)$	$4/\sqrt{2}$	1^a	2^b	$64^{1/5}$	$4^{1/3}$	$(15/\pi)^{1/5}$	$(64/\pi)^{1/3}$
Exponential $(e^{-x},\ x > 0)$	2	2^a	∞^b	∞	$4^{1/3}$	0	$(8/\pi)^{1/3}$
Cauchy $(\pi(1 + x^2))^{-1}$	∞	$2/\pi$	$9/2\pi\sqrt{3}$	∞	∞	∞	∞
$1/(2\sqrt{x}),\ 0 \le x \le 1$	$4/3\sqrt{2}$	∞^a	∞^b	∞	∞	0	0
Student's $t3$ $\left(\frac{2/\pi\sqrt{3}}{(1 + x^2/3)^2}\right)$	$\sqrt{2\pi}\,3^{1/4}$	$4/(\pi\sqrt{3})$	$500/81\pi\sqrt{5}$	$(2000\pi/54\sqrt{5})^{1/5}$	$4^{1/3}$	$(81\pi 3^{3/4}/100)^{2/5}$	$(\pi 3^{3/4})^{2/3}$

[a] Limit for a sequence of densities in $\mathcal{F}_H$.

[b] Generalized definition of $\int\lvert f''\rvert$.

absurd). For the normal density, there is a vast literature on robust scale parameter estimates with a sophisticated underlying theory. For example, one estimate that has received some attention for estimating σ in the normal (μ, σ^2) density is

$$\hat{\sigma} = 0.1174\big(X_{(0.9765n)} - X_{(0.0235n)} + 2\big(X_{(0.8721n)} - X_{(0.1279n)}\big)\big)$$

(Kulldorf, 1963, 1964).

Table 1 contains more valuable information. We have established in Chapters 4 and 5 that $B^*(f)$ and $B_H(f)$ measure the difficulty involved in estimating f by the kernel and histogram methods, respectively. For the histogram estimate we know that the uniform $[0, 1]$ density is the easiest f (B_H attains its minimal value 1). Thus, discontinuities offer no problems for the histogram estimate, as long as f remains bounded. For unbounded f such as $1/2\sqrt{x}$ we have $B_H(f) = \infty$ because $\int |f'| = \infty$. Of course, neither the histogram estimate nor the kernel estimate can handle long-tailed densities very well (see, e.g., the Cauchy density, for which $B^*(f) = B_H(f) = \infty$). Note also the difference between the exponential density and the double exponential density for the kernel estimate, due to the fact that the discontinuity at 0 is absent in the double exponential density. Similarly, $B^*(f) = \infty$ for the uniform $[0, 1]$ density. The best density for the kernel estimate is the isosceles triangular density.

Table 1 is slightly misleading. One is led to believe that with the suggested choices for h, the uniform $[0, 1]$ density, for example, is better estimated by the histogram estimate. This is not true: the table just *suggests* that the optimal $n^{-2/5}$ rate for $E(J_n)$ cannot be achieved for the kernel estimate. It is a good exercise to show that for the uniform $[0, 1]$ density, the best rate achievable for $E(J_n)$ with the kernel estimate is $n^{-1/3}$, and that it is attained if we let h vary as $c/n^{1/3}$ (see Section 7 below).

All the densities in Table 1 are unimodal. It is intuitively obvious that multimodal densities are more difficult to estimate. To study the influence of several peaks on B^* and B_H, consider a central density f with support on $[0, 1]$ and define the following multipeaked density

$$g(x) = \frac{1}{N} \sum_{i=1}^{N} f(x - 2i).$$

Since $\int \sqrt{g} = \sqrt{N} \int \sqrt{f}$, $\int |g'| = \int |f'|$ and $\int |g''| = \int |f''|$, we have

$$B^*(g) = N^{2/5} B^*(f)$$

and

$$B_H(g) = N^{1/3} B_H(f).$$

The presence of several peaks seems to influence the kernel estimate worse than it does the histogram. The parametric method outlined above should be used with extreme care for multimodal densities. A good strategy is to isolate the different peaks and cut the estimation problem into several (easier, unimodal) subproblems.

We have seen that for the kernel estimate, the optimal choice for h is of the form $h = cn^{-1/5}$ when the density f is sufficiently well behaved. It is important to know how the performance, that is, $E(J_n)$, deteriorates when c is suboptimal. Such a study is usually called a *sensitivity analysis*.

The exact expression of $J(n, h)$ given in Theorem 1 is too difficult to handle analytically, but the upper bound for $J(n, h)$ is $n^{-2/5}(ac^{-1/2} + bc^2)$, where

$$a = \sqrt{\frac{2}{\pi}}\,\alpha \int \sqrt{f}, \qquad b = \frac{\beta}{2} \sup_{u>0} \int |(f * \phi_u)''|.$$

It is easy to verify that this upper bound considered as a function of c is minimal when $c = (a/4b)^{2/5}$, and that the upper bound becomes $n^{-2/5}$ times $C^*A(K)B^*(f)$ in that case. Let us call this optimal c c_0. When the actual c is equal to rc_0, then the upper bound for $J(n, h)$ becomes $n^{-2/5}$ times $B^*(f)A(K)C^*$ times $H(r)$ where $H(r) = \frac{1}{5}r^2 + 4/5\sqrt{r}$. When $c = c_0/r$, the extra factor in the upper bound for $J(n, h)$ is $G(r) = 1/5r^2 + \frac{4}{5}\sqrt{r}$. Both $H(r)$ and $G(r)$ are of course minimal and equal to 1 when $r = 1$. We will now show that for all $r > 1$, $G(r) < H(r)$: thus, if we must overestimate c_0 by a factor of r, or underestimate it by a factor of r, it is better to underestimate c_0.

Proof of $G(r) < H(r)$. We need to show that for $r > 1$, $4r^{3/2} + r^4 > 4r^{5/2} + 1$. Set $r = 1 + u$, $u > 0$, and note that by a truncated Taylor series expansion,

$$\begin{aligned}
4r^{3/2} + r^4 &\geq 4\left(1 + \frac{3}{2}u\right) + (1 + 6u + 10u^2 + 6u^3 + u^4) \\
&= 5 + 12u + 10u^2 + 6u^3 + u^4 > 5 + 10u + \frac{15}{2}u^2 + \frac{15}{2}u^3 \\
&= 1 + 4\left(1 + \frac{5}{2}u + \left(\frac{5}{2}\cdot\frac{3}{2}\right)\frac{u^2}{2!} + \left(\frac{5}{2}\cdot\frac{3}{2}\cdot\frac{1}{2}\right)\frac{u^3}{3!}\right) \\
&\geq 1 + 4(1 + u)^{5/2} = 1 + 4r^{5/2}.
\end{aligned}$$

We finally note that the ratio $H(r)/G(r)$ varies as follows: 1 ($r = 1$), $1.038\cdots$ ($r = 1.21$), $1.156\cdots$ ($r = 2$), $2.232\cdots$ ($r = 4$), and $6.787\cdots$ ($r = 9$).

To close this section, we have one interesting observation regarding the optimal values of h as given in (14) and (15), with $\int|f''|$ and $\int|f'|$ replaced by their respective generalized definitions. The observation is that for *all* f,

$$h \le \left(\frac{98415\pi^4}{65536}\right)^{1/5}\sigma n^{-1/5}$$

$$= 6.7726100\cdots\sigma n^{-1/5} \tag{18}$$

for the kernel estimate, and

$$h \le (16\pi^2)^{1/3}\sigma n^{-1/3}$$

$$= 5.40513538\cdots\sigma n^{-1/3} \tag{19}$$

for the histogram estimate, where $\sigma = \sqrt{\mathrm{Var}(X)}$, the standard deviation of f. To see this, use (14) or (15), and combine it with Theorem 5.3 ($B^*(f) \ge (2^9/3^4)^{1/5}$), Theorem 5.9 ($B_H^*(f) \ge 1$, see Section 8 below), and Lemma 5.1 ($(\int\sqrt{f})^2/\sqrt{\mathrm{Var}(X)} \le 2\pi$). These bounds can be used to obtain very rough but useful upper bounds for h in the absence of *any* knowledge about f, if we replace σ by good sample-based estimates.

7. THE UNIFORM DENSITY

The uniform density f on $[0,1]$ warrants separate treatment, because its discontinuities imply that $\liminf_{n\to\infty} n^{2/5}E(J_n) = \infty$ for the kernel estimate. In this section we will show that $E(J_n)$ decreases as $n^{-1/3}$ if h is chosen appropriately. The material in this section could be repeated for densities with more outspoken discontinuities such as the beta densities

$$f(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, \qquad 0 < x < 1, \quad B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

where we can take $a = b \in (0,1]$. It is left as an exercise to show that the optimal rate of convergence to 0 of $E(J_n)$ is $n^{-\varepsilon}$ where ε is a function of a only, and can be chosen arbitrarily close to 0.

In our analysis, a few interesting parallel results should be noted, such as Lemma 18.

LEMMA 18. *For all $p > 0$,*

$$\inf_K \left(\int K^2\right)^p \int |x|^p K \geq \left(\frac{p+1}{2p+1}\right)^p \frac{1}{2p+1},$$

where the infimum is over all densities K on R^1. The infimum is reached for

$$K(x) = \frac{p+1}{2p}(1 - |x|^p), \qquad |x| \leq 1.$$

For $p = 2$, this result is due to Epanechnikov (1969) *and Bartlett* (1963).

Proof. For the density K defined in the statement of Lemma 18, we have $\int K = 1$, $\int |x|^p K = 1/(2p+1)$ and $\int K^2 = (p+1)/(2p+1)$. Because $(\int K^2)^p \int |x|^p K$ is scale invariant, it suffices to take the infimum over all densities K with $\int |x|^p K = 1/(2p+1)$. All the densities considered here are normalized in this manner. Any density g can be written as $g = K + g^*$ where $\int g^* = 0$ and $\int |x|^p g^* = 0$ (because $\int |x|^p g = \int |x|^p K$). Thus,

$$\begin{aligned}
\int g^2 &= \int K^2 + \int g^{*2} + 2\int_{-1}^{1} \frac{p+1}{2p}(1 - |x|^p) g^* \\
&= \int K^2 + \int g^{*2} + \frac{p+1}{p} \int_{[-1,1]^c} (|x|^p - 1) g^* \\
&\geq \int K^2 + \int g^{*2},
\end{aligned}$$

because $g^* \geq 0$ on $[-1,1]^c$ (otherwise, g would not be a density), and $|x|^p \geq 1$ on $[-1,1]^c$. The right-hand side of the inequality is minimal if $g^* = 0$ almost everywhere. This concludes the proof of Lemma 18.

LEMMA 19. *Let $B_n = E(f_n) - f$. Then, for the kernel estimate* (1), *and all $K \in L_1$,*

$$E\left(\int |f_n - f|\right) \geq \int |B_n| \sim 2h \int |x| K(x)\, dx \quad \text{as } h \to 0.$$

Proof. The first inequality follows from Jensen's bound. For the last part, we note that $E(f_n)$ is smaller than or equal to one on $[0,1]$, and thus that $\int |B_n| = 2\int_{[0,1]^c} E(f_n)$.

But if K has distribution function F, and Y is a random variable with density K,

$$\begin{aligned}\int_{[0,1]^c} E(f_n) &= \int_{[0,1]^c} F\left(\frac{1-x}{h}\right) - F\left(-\frac{x}{h}\right) dx \\ &= \int_{-\infty}^{0} (-P(hY > 1-x) + P(hY > -x))\, dx \\ &\quad + \int_{1}^{\infty} (P(hY < 1-x) - P(hY < -x))\, dx \\ &= \int_{0}^{\infty} (P(hY < -x) + P(hY > x))\, dx \\ &\quad + \int_{1}^{\infty} (-P(hY > x) - P(hY < -x))\, dx \\ &= hE(|Y|) + o(h)\end{aligned}$$

if $E(|Y|) < \infty$ and $h \to 0$.

LEMMA 20. *If $h \to 0$, $nh \to \infty$, and K is a bounded density with compact support, we have $E(J_n) \geq (\alpha + o(1))\sqrt{2/\pi nh}$, where $\alpha = \sqrt{\int K^2}$.*

Proof. Lemma 9 and the inequality $\psi(u) \geq \sqrt{2/\pi}$ imply

$$E(J_n) \geq \int_T \sqrt{\frac{2}{\pi}}\, \sigma_n - \frac{cK^*\lambda(T^*)}{nh},$$

where $T = [0,1]$. Also, $\lambda(T^*) \to 1$ because $h \to 0$. By Lemma 10,

$$\int_T \sigma_n \sim \frac{\alpha}{\sqrt{nh}} \int \sqrt{f} = \frac{\alpha}{\sqrt{nh}}.$$

This concludes the proof of Lemma 20.

THEOREM 7. *If K is a bounded density with compact support, we have for kernel estimate* (1) *applied to the uniform density f on* $[0,1]$,

$$\liminf_{n\to\infty} \inf_{h>0} n^{1/3} E(J_n) \geq \left(\frac{4}{\pi} \int K^2 \int |x| K\right)^{1/3} \geq \left(\frac{8}{9\pi}\right)^{1/3}.$$

Proof. By mimicking the proof of Theorem 2, we have

$$\inf_h E(J_n) \geq \max_{t>0} \min\left(2t \int |x| K, \sqrt{\frac{2\alpha}{\pi nt}}\right)(1 + o(1))$$

where we applied Lemmas 19 and 20. The maximum is attained for $t = (\alpha^2/(2\pi n(\int|x|K)^2))^{1/3}$, and resubstitution gives the middle expression in Theorem 7. The K-independent lower bound is obtained by applying Lemma 18 with $p = 1$.

THEOREM 8. *If K is a bounded density with compact support, then the kernel estimate* (1) *applied to the uniform density on* $[0,1]$ *satisfies*

$$\limsup_{n\to\infty} \inf_{h>0} n^{1/3}E(J_n) \le \left(\left(\frac{8}{\pi}\right)^{1/3} + \left(\frac{1}{\pi}\right)^{1/3}\right)\left(\int K^2 \int |x|K\right)^{1/3}.$$

Proof. Let $T = [0,1]$, and let all the undefined symbols be as in Sections 2 and 3. From the proof of Theorem 1,

$$E(J_n(T)) = \int_T \sigma_n \psi\left(\frac{|B_n|}{\sigma_n}\right) + O\left(\frac{1}{nh}\right).$$

By Lemma 10,

$$\int_T \sigma_n = (\alpha + o(1))\int \frac{\sqrt{f}}{\sqrt{nh}} = \frac{\alpha + o(1)}{\sqrt{nh}}.$$

Now, by $\psi(u) \le u + \sqrt{2/\pi}$ and Lemma 19,

$$\begin{aligned} E(J_n) &= E(J_n(T^c)) + E(J_n(T)) \\ &= E\left(\int_{T^c} f_n\right) + E(J_n(T)) \\ &\le \int_{T^c} E(f_n) + \sqrt{\frac{2}{\pi}} \int_T \sigma_n + \int_T |B_n| \\ &= 2h\int |x|K + o(h) + \sqrt{\frac{2}{\pi}}\,\frac{\alpha}{\sqrt{nh}} + o\left(\frac{1}{\sqrt{nh}}\right). \end{aligned}$$

The main terms in this upper bound are minimized by taking $h^{3/2} = (\alpha\sqrt{2/\pi})/(4\int|x|K\sqrt{n})$. We obtain

$$n^{1/3}E(J_n) \le \left(\frac{2}{\pi}\int|x|K\int K^2\right)^{1/3}(4^{1/3} + 2^{-1/3} + o(1)),$$

which was to be shown.

We did not derive the exact optimal asymptotic behavior for the kernel estimate. Yet, by relatively sloppy arguments, we obtained an upper bound (Theorem 8) and a lower bound (Theorem 7) with a ratio

$$2^{1/3} + 2^{-2/3} = 1.8898816\cdots,$$

valid for all K that are bounded and have compact support. The value of h for which the upper bound of Theorem 8 is attained is

$$h = \left(\frac{\int K^2}{\left(\int |x| K\right)^2 8\pi n} \right)^{1/3}.$$

Thus, if not enough information is available to exclude the possibility that f is the uniform $[0, 1]$ density, it is dangerous to let h vary as $c^* n^{-1/5}$. Indeed, from Lemma 19 we see that for the uniform density

$$E(J_n) \geq \left(\int |x| K + o(1) \right) 2c^* n^{-1/5},$$

a rate that is well above the optimal rate given in Theorems 7 and 8.

Let us finally note that for the uniform density f, the optimal kernel is not the Epanechnikov kernel (which is optimal for the restricted class $\mathscr{F}$ of Theorem 1), but rather the isosceles triangular density $1 - |x|$, $|x| \leq 1$. This follows from Theorems 7 and 8 and Lemma 18. For other members of the beta family, the optimal kernel is different: its shape depends upon the kind of discontinuity that occurs.

8. A MINIMAX STRATEGY FOR CHOOSING THE SMOOTHING FACTOR

There are situations in which one is uncertain about the smoothness of f, for example, when one suspects that f has a discontinuity. In such cases, Theorem 1 gives us no clue as to how h should be chosen for the kernel estimate. In fact, as we have seen in Section 7, when f is the uniform $[0, 1]$ density, it is outright dangerous to choose h as a constant times $n^{-1/5}$. We could play a conservative game by enlarging the class of densities, deriving an upper bound for $E(J_n)$ for this class, and obtaining h by minimizing this upper bound. This is a minimax strategy of sorts. Wertz (1972) developed a similar strategy for L_2 properties of the kernel estimate.

Consider first the class of densities $\mathscr{F}$ of Section 4, that is, all absolutely continuous densities f with bounded and continuous a.e. derivative f', and thus $\int |f'| < \infty$. For this class, we defined the following factor:

$$B_H(f) = \left(\frac{1}{2} \left(\int \sqrt{f} \right)^2 \int |f'| \right)^{1/3}.$$

We will slightly generalize this definition by defining

$$B_H^*(f) = \left(\frac{1}{2} \left(\int \sqrt{f} \right)^2 \sup_{a>0} \int |(f * \phi_a)'| \right)^{1/3},$$

where $\phi \in \mathscr{F}$ is a continuously differentiable density with compact support.

The main message in this section is that if h is chosen as a constant times $n^{-1/3}$, then $E(J_n) = O(n^{-1/3})$ for all densities with compact support and finite $B_H^*(f)$. This class includes all absolutely continuous densities with compact support, and even densities with simple discontinuities such as the uniform $[0, 1]$ density. We will derive some properties in parallel with the derivations of Sections 2 and 3. For some lemmas, we only sketch the proofs.

LEMMA 21. *Let f and K be arbitrary densities on R. Then, when $\gamma = \int |x| K$,*

$$\int |f * K_h - f| \le h\gamma \liminf_{a \downarrow 0} \int |(f * \phi_a)'|, \quad \textit{all } h > 0.$$

Proof. First consider $f \in \mathscr{F}$. Since

$$f(y) - f(x) = \int_x^y f'(z)\, dz,$$

we have

$$\int |f * K_h - f| = \int \left| \int \tilde{K}\left(\frac{x-y}{h} \right) f'(y)\, dy \right| dx,$$

where

$$\tilde{K}(x) = \begin{cases} \displaystyle\int_x^\infty K(z)\, dz, & x \ge 0, \\ \displaystyle\int_{-\infty}^x K(z)\, dz, & x < 0. \end{cases}$$

This implies that

$$\int |f * K_h - f| \le h \int |f'| \int |\tilde{K}| = h \int f |f'| \gamma.$$

The extension to all f is as in the proof of Lemma 4 (ii) and is based upon the fact that $f * \phi_a$ is in $\mathscr{F}$ for all f and all $a > 0$.

THEOREM 9. *For all $f \in \mathscr{F}$, $B_H^*(f) = B_H(f)$. For all f, $B_H^*(f) \ge 1$, and this bound is attained for the uniform density on* $[0,1]$.

Proof. The first statement is not very hard to show (see, e.g., Lemma 2 for a similar proof) and is not proved here. The second statement is partially shown in Lemma 15, that is, for all $f \in \mathscr{F}$. Here we will prove that it holds for all f. We will use the inequalities $\int \sqrt{f} \ge 1/\sqrt{\sup f}$ and $\int |(f * \phi_a)'| \ge 2 \sup(f * \phi_a)$ (see Lemma 15): this gives

$$B_H^*(f)^3 \ge \frac{\sup(f * \phi_a)}{\sup f}, \quad \text{all } a > 0.$$

But since $f * \phi_a \to f$ for almost all x as $a \downarrow 0$ (Theorem 2.3), it is clear that $\sup_{a>0} \sup(f * \phi_a) \ge \sup f$ (incidentally, we always have $\sup(f * \phi_a) \le \sup f$), and thus $B_H^*(f)^3 \ge 1$.

THEOREM 10. *Let K be a bounded density with compact support, and let h satisfy* (3). *Then for the kernel estimate* (1) *and all f with compact support,*

$$E(J_n) \le \sqrt{\frac{2}{\pi}} \frac{\alpha \int \sqrt{f}}{\sqrt{nh}} + h\gamma \sup_{a>0} \int |(f * \phi_a)'| + o\left((nh)^{-1/2}\right).$$

Furthermore,

$$\limsup_{n \to \infty} \inf_h n^{1/3} E(J_n) \le C_1^* A_1(K) B_H^*(f),$$

where

$$C_1^* = \frac{3}{\pi^{1/3}} = 2.0483522 \cdots$$

and

$$A_1(K) = (\alpha^2\gamma)^{1/3}.$$

The upper bound $C_1^* A_1(K) B_H^*(f)$ is not exceeded if $B_H^*(f)$ is finite and h is taken as follows:

$$h = \left[\sqrt{\frac{2}{\pi}} \frac{\alpha \int \sqrt{f}}{2\gamma \sup_{a>0} \int |(f * \phi_a)'|} \right]^{2/3} n^{-1/3}.$$

Proof. As in the proof of Theorem 1, we have, for bounded intervals T,

$$\int_T E(|f_n - f|) \le \int_T \left(\sqrt{\frac{2}{\pi}}\, \sigma_n + |B_n| \right) + \frac{cK^*}{nh} \lambda(T)$$

$$\le \sqrt{\frac{2}{\pi}} \frac{\alpha}{\sqrt{nh}} \int \sqrt{f} + h\gamma \sup_{a>0} \int |(f * \phi_a)'| + o\left((nh)^{-1/2}\right),$$

where we used Lemmas 10 and 21. If K has compact support, and we take T large enough, then $E(J_n) = E(J_n(T))$ for all n, and we are done.

The main terms in the upper bound are of the form $uh^{-1/2} + vh$. Considered as a function of h, this is minimal when $h = (u/2v)^{2/3}$, and the minimal value is $(u^2 v)^{1/3} 3/4^{1/3}$. But this can be rewritten as $C_1^* A_1(K) B_H^*(f)$, and the Theorem is proved.

A few remarks are in order here. First, by Lemma 18, $A_1(K)$ is at least equal to $(\frac{2}{9})^{1/3}$, and this minimum is attained for the isosceles triangular density $1 - |x|$ on $[-1, 1]$. Resubstitution of this value for $A_1(K)$ in the upper bound gives $(6/\pi)^{1/3} B_H^*(f) = 1.240701 \cdots B_H^*(f)$. It is better than the upper bound of Theorem 6 for the histogram estimate, and very close to the lower bound of Theorem 5 for the histogram estimate. The value of h suggested in the Theorem is

$$h = \left(\frac{3}{\pi} \right)^{1/3} \left(\frac{\int \sqrt{f}}{\sup_{a>0} \int |(f * \phi_a)'|} \right)^{2/3} n^{-1/3}$$

when the optimal k is used (just substitute $\alpha = \sqrt{2/3}$ and $\gamma = 1/3$ in the formula for h).

Finally, because the upper bound of Theorem 10 is minimal for the uniform density on $[0,1]$ (see Theorem 9), it is to the advantage of the user to "uniformize" the data as much as possible by transformations prior to constructing the kernel estimate.

9. LIPSCHITZ CLASSES, BRETAGNOLLE–HUBER CLASSES, AND UNIFORM UPPER BOUNDS

As in Section 4.2, we call $W(s,\alpha,C)$ the *Lipschitz class* with parameters s,α,C, that is, the class of all densities on $[0,1]$ with $(s-1)$ absolutely continuous derivatives for which for all $x, y \in R$,

$$|f^{(s)}(x) - f^{(s)}(y)| \le C|x-y|^{\alpha}.$$

Here $\alpha \in (0,1]$, s is a nonnegative integer, and $C > 0$. We will call $F_{s,r}$ the *Bretagnolle–Huber class* with parameters s and r (s is a positive integer and $r > 0$), that is, the class of all densities on $[0,1]$ for which $D_s^*(f) \le r$, where $D_s^*(f)$ is defined as follows:

$$D_s^*(f) = \left(\left(\int \sqrt{f} \right)^{2s} \sup_{a>0} \int \left| (f * \phi_a)^{(s)} \right| \right)^{1/(2s+1)}.$$

Here ϕ is an even bounded density, monotone on $[0,\infty)$, having s absolutely continuous derivatives and compact support. Note that this definition is slightly different from the definitions of similar quantities in Chapter 4 and Sections 5.1–5.8.

In Theorem 4.6 we have seen that for C large enough and for all density estimates f_n,

$$\sup_{f \in W(s,\alpha,C)} E\left(\int |f_n - f| \right) \ge (c_3 + o(1)) C^{1/(2(s+\alpha)+1)} n^{-(s+\alpha)/(2(s+\alpha)+1)},$$

in the notation of Theorem 4.6. Recall that $c_3 = c_3(s,\alpha) > 0$. In Theorem 4.3, a similar minimax lower bound was obtained for $F_{s,r}$: $c_4(r)n^{-s/(2s+1)}$, valid for all r large enough.

What we would like to do now is to show that these minimax bounds can be achieved by the kernel estimate up to a proportionality constant *not*

depending upon n, C, or r, at least if we allow a slightly more general definition of the kernel estimate. The importance of this fact should be stressed very strongly. Other estimators can only at best reduce a proportionality constant by L_1 minimax standards. Thus, only for special classes of densities (such as monotone densities on $[0,1]$, etc.) should we seriously consider other estimators.

For Lipschitz classes, we will only consider the case $\alpha = 1$, the case $0 < \alpha < 1$ being less interesting anyway. It is a good exercise nevertheless to treat that case after having seen the general treatment for $\alpha = 1$. Also, we will not take the long route: upper bounds will be obtained very simply by separating the bias and variance terms rather crudely. The only effect of this is that the proportionality constants are slightly worse. In our treatment, we will follow to some extent Bretagnolle and Huber (1979).

LEMMA 22 (Uniform Bounds for the Bias). *Let K be a measurable function satisfying*:

$$K \text{ is symmetric}, \quad \int K = 1, \int x^i K = 0, \ i = 1, \ldots, s-1,$$

$$\int |x|^s |K| < \infty,$$

and let L be the kernel associated with K, *that is*,

$$L(x) = (-1)^s \int_x^\infty \frac{(y-x)^{s-1}}{(s-1)!} K(y)\, dy, \qquad x > 0,$$

$$L(-x) = -(-1)^s L(x), \qquad x < 0.$$

Then $\int |L| < \infty$. For $s = 1$, $K \geq 0$, $\int |L| = \int |x| K$, and for $s = 2$, $K \geq 0$, $\int |L| = \int x^2 K/2$. If f has $(s-1)$ absolutely continuous derivatives, then

$$\int |f * K_h - f| \leq h^s \int |L| \int |f^{(s)}|.$$

For all f,

$$\int |f * K_h - f| \leq h^s \int |L| \liminf_{a \downarrow 0} \int |(f * \phi_a)^{(s)}|.$$

When $f \in W(s-1, 1, C)$, $s \geq 1$, then the latter upper bound is not greater

than

$$Ch^s \int |L|.$$

Proof.

$$\int |L| = 2\int_0^\infty \left| \int_x^\infty \frac{(y-x)^{s-1}}{(s-1)!} K(y)\, dy \right| dx$$

$$\le 2\int_0^\infty \left(\int_0^y \frac{(y-x)^{s-1}}{(s-1)!}\, dx \right) |K(y)|\, dy = 2\int_0^\infty \frac{|y|^s}{s!} |K(y)|\, dy < \infty.$$

For $s = 1$, $s = 2$, and $K \ge 0$, we obtain equality throughout, which gives us simple explicit expressions for $\int |L|$.

When f has $(s-1)$ absolutely continuous derivatives, then, by Taylor's series expansion,

$$f(x+y) - f(x) = \sum_{j=1}^{s-1} \frac{y^j}{j!} f^{(j)}(x) + \int_x^{x+y} \frac{(x+y-u)^{s-1}}{(s-1)!} f^{(s)}(u)\, du.$$

If L_h is defined as $(1/h)L(x/h)$, and $(L)_h$ is defined as L with K_h instead of K, then

$$f * K_h - f = f^{(s)} * (L)_h = h^s f^{(s)} * L_h.$$

Thus,

$$\int |f * K_h - f| \le h^s \int |L_h| \int |f^{(s)}| = h^s \int |L| \int |f^{(s)}|.$$

For any f and fixed $h > 0$, we know that for almost all x,

$$|f * K_h - f| = \liminf_{a \downarrow 0} |(f * K_h - f) * \phi_a| = \liminf_{a \downarrow 0} |f * \phi_a * K_h - f * \phi_a|.$$

Thus, by Fatou's lemma, we have for the same h,

$$\int |f * K_h - f| \le \liminf_{a \downarrow 0} \int |f * \phi_a * K_h - f * \phi_a|$$

$$\le \liminf_{a \downarrow 0} h^s \int |(f * \phi_a)^{(s)}| \int |L|.$$

We must now bound the integral in the last expression from above by C for all f in $W(s-1,1,C)$: $\int |f^{(s-1)} * \phi_a'| \le C$. To prove this, note that $f^{(s-1)}$ is Lipschitz (C). Now, for any Lipschitz (C) function g on R, we have

$$\begin{aligned}
\left|\int g(y)\phi_a'(x-y)\,dy\right| &= \left|\int (g(y)-g(x))\phi_a'(x-y)\,dy\right| \\
&\le \int |g(y)-g(x)||\phi_a'(x-y)|\,dy \\
&\le C\int |x-y||\phi_a'(x\text{-}y)|\,dy \\
&= C\int |z||\phi'(z)|\,dz \\
&= -2\int_0^\infty Cz\phi'(z)\,dz \\
&= 2C\int_0^\infty \phi(z)\,dz \\
&= C.
\end{aligned}$$

This concludes the proof of Lemma 22, because $\liminf_{a\downarrow 0}\int_{[0,1]^c}|g*\phi_a'| = 0$ for all Lipschitz (C) functions g vanishing outside $[0,1]$. (To see this, note that $g*\phi_a'$ is absolutely bounded by C, and is zero outside $[-aM, 1+aM]$ where M is a number depending upon the support of ϕ.)

LEMMA 23 (Uniform Bounds for the Variation). *Let K be a kernel on R satisfying the conditions of Lemma* 22, *let K vanish outside* $[-1,1]$, *and define* $C_s = \liminf_{a\downarrow 0}\int |(f*\phi_a)^{(s)}|$ *and* $\alpha = \sqrt{\int K^2} < \infty$. *Then, if f vanishes outside* $[0,1]$,

$$E\left(\int |f_n - f*K_h|\right) \le (nh)^{-1/2}\left(\alpha\int\sqrt{f} + \sqrt{C_1(h+2h^2)\int |x|K^2/2}\right).$$

Proof. By the Cauchy–Schwarz inequality applied to $E(|f_n - f*K_h|)$, we obtain the inequality

$$E\left(\int |f_n - f*K_h|\right) \le (nh)^{-1/2}\int\sqrt{f*K_h^2},$$

where $K_h^2(x) = (1/h)K^2(x/h)$. If we introduce $K^\dagger = K^2/\int K^2$, then the upper bound can be rewritten as

$$(nh)^{-1/2}\alpha\int\sqrt{f * K_h^\dagger} \le (nh)^{-1/2}\alpha\left(\int\sqrt{f} + \int\sqrt{|f - f * K_h^\dagger|}\right)$$

$$\le (nh)^{-1/2}\alpha\left(\int\sqrt{f} + \sqrt{\int_{-h}^{1+h} dx \int |f - f * K_h^\dagger|}\right)$$

(by Cauchy's inequality)

$$\le (nh)^{-1/2}\alpha\left(\int\sqrt{f} + \sqrt{(1+2h)hC_1\int |x|K^\dagger/2}\right)$$

(by Lemma 22),

which was to be shown.

THEOREM 11 (Minimax Upper Bounds). *Let K and L be as in Lemmas 22 and 23, and let α and C_s be as defined in Lemma 23. Then, for all f vanishing outside $[0,1]$, and the kernel estimate f_n with kernel K,*

$$E\left(\int |f_n - f|\right)$$

$$\le h^s C_s \int |L| + (nh)^{-1/2}\left(\alpha\int\sqrt{f} + \sqrt{C_1(h+2h^2)\int |x|K^2/2}\right).$$

In particular, we have the following minimax upper bound for $W(s-1,1,C)$, all $s \ge 1$: if g_n denotes any density estimate,

$$\inf_{g_n} \sup_{f \in W(s-1,1,C)} E\left(\int |g_n - f|\right)$$

$$\le \inf_{\substack{h>0 \\ K}} \sup_{f \in W(s-1,1,C)} E\left(\int |f_n - f|\right)$$

$$\le \inf_K \frac{2s+1}{2s}\left(2sC\int |L|\alpha^{2s} n^{-s}\right)^{1/(2s+1)} (1 + o(1)).$$

It is understood that the infima are taken over all K satisfying the conditions of

Lemmas 22 and 23. The last inequality is obtained, for example, by choosing

$$h = \left(\frac{\alpha}{2sC\int |L|} \right)^{2/(2s+1)} n^{-1/(2s+1)}.$$

In particular, for $s = 1$, the upper bound is $(1 + o(1))$ times

$$\inf_{K \geq 0} \frac{3}{2} \left(2C \int |x| K \alpha^2 n^{-1} \right)^{1/3} = \left(\frac{3C}{2n} \right)^{1/3},$$

after having taken the triangular kernel K (with $\alpha^2 = \frac{2}{3}$, $\int |x| K = \frac{1}{3}$). For the case $s = 2$, the infimum of the upper bound is reached for Epanechnikov's kernel $\frac{3}{4}(1 - x^2)_+$ (this has $\int x^2 K = \frac{1}{5}$ and $\int K^2 = \frac{3}{5}$), and reads

$$(1 + o(1))\left(\tfrac{225}{512} C\right)^{1/5} n^{-2/5}.$$

Proof. The first inequality follows from Lemmas 22 and 23. The second chain of inequalities requires three facts: first, for all f vanishing outside $[0, 1]$, we have $C_1 \leq C_2 \leq C_3 \leq \cdots \leq C_s$. Also, for any $f \in W(s - 1, 1, C)$, $C_s \leq C$ (see proof of Lemma 22). Second, $\int \sqrt{f} \leq 1$ when f is zero outside $[0, 1]$. Third, the function $h^s C \int |L| + \alpha / \sqrt{nh}$ is minimal for the choice of h given in the theorem. Formal replacement of this value of h gives the upper bound. The remainder of the proof is trivial.

The minimax lower bound (Theorem 4.6) for $W(s - 1, 1, C)$ and the minimax upper bound of Theorem 11 have the same dependency upon C and n. They differ only in a proportionality constant, which in turn depends only upon s. It is informative to know what the gap is between the bounds for the most important classes, $W(0, 1, C)$ and $W(1, 1, C)$: From Theorem 4.7, we recall that for $W(0, 1, C)$, the coefficient of $(C/n)^{1/3}$ is $\frac{21}{160}(\frac{12}{25})^{1/3}$, so that the ratio between upper and lower bound is about 11. For $W(1, 1, C)$, this ratio is of the order of 30. The upper bounds are without any doubt very loose: they are obtained for a rather primitive estimator, the kernel estimate with smoothing factor h chosen as a function of s, C, and K only! It seems of course much more efficient to choose h as a function of f (see Chapter 6 for the automatic choice of h), but if everything else fails, or at least as a rough first guess, one can take the pessimistic attitude that the minimax error should be minimized for a certain class such as $W(s - 1, 1, C)$. In that case, h can be chosen as indicated in Theorem 11.

In addition to h, we should choose K. For the cases $s = 1$ and $s = 2$, we know that K should be the isosceles triangular density on $[-1, 1]$ and

Epanechnikov's kernel. For $s \geq 3$, K must necessarily take negative values, and f_n may no longer be a density because of this, although its integral is still one. However, these estimates can easily be normalized, as shown in Section 7.6. Bartlett (1963) was the first person to indicate that better rates of convergence can be obtained by taking kernels such as those of Lemmas 22 and 23 (see also Section 7.6 for a detailed treatment). For $s = 4$, he obtained the optimal form of K too. In general, kernels K satisfying the conditions of Lemmas 22 and 23 can be constructed without great difficulty. For example, start with a basic symmetric density K vanishing outside $[-1, 1]$. For fixed s as in Lemma 22, we need only find real numbers p_i (not necessarily positive) such that the function

$$\sum_{i=1}^{N} p_i K_{1/i}$$

will do. (Incidentally, the choice $1/i$ is arbitrary and can be replaced by other positive numbers taken from $(0, 1)$.) For example, when K is uniform on $[-1, 1]$, this gives conditions of the following type:

$$\sum p_i = 1; \quad \sum p_i i^{-2} = 0; \quad \sum p_i i^{-4} = 0; \ldots; \sum p_i i^{-(s-2)} = 0,$$

for s even, $s \geq 4$. Generally, there is a solution with $N = s/2$ components in the mixture. See also Bretagnolle and Huber (1979) for other constructions, based upon Legendre polynomials.

From the first inequality of Theorem 11, we see that for individual f, the upper bound for the expected L_1 error can be much smaller than the minimax upper bound. For example, $\int \sqrt{f}$ is very crudely bounded from above by 1, although we have the following fact:

LEMMA 24. *For all f in $W(0, 1, C)$, $C \geq 4$, the following inequality is valid*:

$$\frac{4}{3\sqrt{2}} \geq \int \sqrt{f} \geq \frac{4}{3C^{1/4}}.$$

Both inequalities can be achieved.

Proof. The upper bound is achieved for the isosceles triangular density on $[0, 1]$ (increasing as $4x$ on $[0, \frac{1}{2}]$. (It is in a sense the "smoothest" density in $W(0, 1, C)$.) The lower bound is achieved by the isosceles triangular density on $[\frac{1}{2} - b, \frac{1}{2} + b]$, where $b = 1/\sqrt{C}$ (the slope of the edges is C of course).

For large C, we see that the minimax upper bound of Theorem 11 loses some of its power, because it does not provide us with good information about most densities contained in the given Lipschitz class.

To obtain a uniform upper bound for the Bretagnolle–Huber classes $F_{s,r}$, the approach taken for $W(s,1,C)$ cannot be followed, simply because the h that would give us the right dependency upon r is not a function of s, r, and n only, but also of C_s and $\int\sqrt{f}$. But since C_s and $\int\sqrt{f}$ are not known, they must be estimated. Thus, strictly speaking, we should only consider kernel estimates in which h is chosen as a function of the data in such a way that, asymptotically, h approaches the optimal h. This adaptive strategy was followed by Bretagnolle and Huber (1979) in their quest for a minimax upper bound for $F_{s,r}$. Note also that, on $F_{s,r}$, $\int\sqrt{f}$ is not uniformly bounded from below, and that C_s is not uniformly bounded from above. (This follows by using different scales for the same density!) Since the minimax upper bound for $F_{s,r}$ requires the following tedious work, it will not be proved here: first, cut the data into pieces, and use one of the small pieces ($o(n)$ in size) to estimate h, and use the big piece (of size $\sim n$) to construct the kernel estimate with this h. Then, $\sup_{f\in F_{s,r}} E(\int|f_n - f|)$ is bounded from above by the expected value of the first expression of Theorem 11, preceded by a supremum over $F_{s,r}$. One must then make sure that this expression is not larger than $rn^{-s/(2s+1)}$ times a constant not depending upon r or n.

The uniform inequality of Theorem 11 has many other uses, besides obtaining minimax upper bounds. For one thing, it is applicable for *all* n, and thus of great value to the person who has to work with a small sample. But more importantly, we can obtain the rate of convergence of a kernel estimate with random smoothing factor h independent of $X_1,\dots,X_n$. Typically, h would be a function of $X_{n+1},\dots,X_{n+m}$. We have the following theorem:

THEOREM 12. *Under the conditions of Theorem* 11, *and with the same notation, we note that for the kernel estimate f_n with smoothing factor h independent of $X_1,\dots,X_n$, the bound of Theorem* 11 *remains valid, provided only that the expected value with respect to h is taken on the right-hand side. In particular, assume that there exists a sequence of positive numbers h_{n0} with $h_{n0}\to 0$, $nh_{n0}\to\infty$, and*

$$E(h^s)\sim h_{n0}^s;\quad E(h^{-1/2})\sim h_{n0}^{-1/2};\qquad E(\sqrt{h})\to 0,$$

and assume that $\int|x|K^2<\infty$. Then

$$E\left(\int|f_n-f|\right)\le\left(h_{n0}^s C_s\int|L|+(nh_{n0})^{-1/2}\alpha\int\sqrt{f}\right)(1+o(1)).$$

If $C_s < \infty$, $\int\sqrt{f} < \infty$, *and*

$$h_{n0} = \left(\frac{\alpha\int\sqrt{f}}{2sC_s\int|L|}\right)^{2/(2s+1)} n^{-1/(2s+1)},$$

then

$$E\left(\int|f_n - f|\right) \le \frac{2s+1}{2s}\left(2s\int|L|\alpha^{2s}n^{-s}\right)^{1/(2s+1)} D_s^*(f)(1+o(1)).$$

For smoothing factors h that depend upon the data, we need a stronger theorem. The consistency of such estimates is dealt with in Chapter 6.

10. DENSITIES WITH UNBOUNDED SUPPORT

We have until now postponed the problem of the performance of density estimates for densities f with unbounded support, and this for two reasons: such cases are less important (data can always be mapped monotonically to $[0,1]$; and densities with unbounded support occur less often in practice), and the additional notational and conceptual burden would only detract from the main ideas.

Lower bounds for all f were obtained in Theorems 2 and 5. Thus, we will content ourselves with the derivation of upper bounds for $E(J_n)$. Only the kernel estimate will be treated here, because the histogram estimate can be treated similarly. If we start from the uniform upper bound of Theorem 11, the proofs become very short. It should be clear though that we are sacrificing a bit with respect to the bounds of Theorems 1 and 6, obtained by exploiting Berry–Esseen type inequalities for deviations from normal behavior. From Lemma 23 and Theorem 11, we obtain the following:

LEMMA 25. *Let f_n be a kernel estimate on R with K and L as defined in Lemmas 22 and 23, and let $s \ge 1$ be an integer. Define*

$$C_s = \sup_{a>0}\int|(f^*\phi_a)^{(s)}|,$$

where ϕ is an even bounded density, monotone on $[0,\infty)$, with s absolutely continuous derivatives and compact support. Then

$$E\left(\int|f_n - f|\right) \le h^sC_s\int|L| + (nh)^{-1/2}\int\sqrt{f^*(K^2)_h}.$$

The uniform bound of Lemma 25 shows the importance of the omnipresent factor $\int\sqrt{f*(K^2)_h}$, to which we will devote a separate lemma.

LEMMA 26 (The Factor $\int\sqrt{f*(K^2)_h}$). *Let K be an arbitrary measurable function on R with conditions added as stated in the various statements of this lemma, and let f be a density on R.*

A. *There exists a density f on R with $\int\sqrt{f} < \infty$, yet $\int\sqrt{f*(K^2)_h} = \infty$ for all h small enough, and for all K with $\int K = 1$, vanishing outside $[-1,1]$, and bounded in absolute value by a constant.*

B. $\int\sqrt{f*(K^2)_h} \geq \int\sqrt{f}\sqrt{\int K^2}$. *In particular,* $\int\sqrt{f*(K^2)_h} = \infty$ *whenever* $\int\sqrt{f} = \infty$.

C. $\int\sqrt{f*(K^2)_h} \to \int\sqrt{f}\sqrt{\int K^2}$ *as* $h \downarrow 0$, *when K has compact support and is bounded, and f satisfies the following conditions:*

(i) *There exist positive numbers t, T, such that*

$$\int_{|x|\geq T}\sqrt{\sup_{|y-x|\leq t} f(y)}\,dx < \infty.$$

(ii) *f is almost everywhere continuous.*

D. *If f and K are both symmetric and unimodal, then*

$$\int\sqrt{f*K^2} \leq 2\int\sqrt{f}\sqrt{\int K^2} + \sqrt{2\int|K|}.$$

Proof. A. Let f be the indicator function of $\cup_{i=1}^{\infty}[2^i, 2^i + 1/i(i+1)]$, $i \geq 1$. Clearly, $\int\sqrt{f} = 1$. Also, for $h \leq \frac{1}{2}$,

$$\begin{aligned}
\int\sqrt{f*(K^2)_h} &\geq \sum_{i=1}^{\infty}\int_{2^i-h}^{2^i+(i(i+1))^{-1}+h}\sqrt{f*(K^2)_h} \\
&\geq \frac{\displaystyle\sum_{i=1}^{\infty}\int_{2^i-h}^{2^i+(i(i+1))^{-1}+h} f*(K^2)_h}{\displaystyle\sup_{-h\leq x-2^i\leq h+(i(i+1))^{-1}}\sqrt{f*(K^2)_h}} \\
&\geq \sum_{i=1}^{\infty}\frac{\sqrt{hi(i+1)}}{K^*} \\
&\quad\cdot\frac{1}{i(i+1)}\int K^2 \qquad \text{(where } K^* \text{ is the uniform bound for } |K|\text{)} \\
&= \infty.
\end{aligned}$$

B. By Jensen's inequality, $f^*(K^2)_h \geq \sqrt{f} * (K^2)_h / \sqrt{\int K^2}$.

C. Let t and T be as in (i), and assume that K vanishes outside $[-1, 1]$. We always have

$$\sqrt{f * (K^2)_h(x)} \leq \sqrt{\sup_{|y-x| \leq h} f(y) \int K^2}.$$

The right-hand side of this inequality is smaller than a fixed integrable function for $h \leq t$ (by (i)). By Theorem 2.3, $f * (K^2)_h \to f \int K^2$ for almost all x, so that, by the Lebesgue dominated convergence theorem,

$$\int_{[-T,T]^c} \sqrt{f * (K^2)_h} \to \int_{[-T,T]^c} \sqrt{f} \sqrt{\int K^2}.$$

Also, if we set $K^\dagger = K^2 / \int K^2$, then

$$\begin{aligned}
\int_{-T}^{T} \left| \sqrt{f * (K^2)_h} - \sqrt{f} \sqrt{\int K^2} \right| &\leq \sqrt{\int K^2} \int_{-T}^{T} |\sqrt{f * K_h^\dagger} - \sqrt{f}| \\
&\leq \sqrt{\int K^2} \int_{-T}^{T} \sqrt{|f * K_h^\dagger - f|} \\
&\leq \sqrt{\int K^2} \sqrt{2T \int |f * K_h^\dagger - f|} \\
&= o(1),
\end{aligned}$$

where we used Theorem 2.1.

D. Assume that $x \geq 0$. Then

$$\begin{aligned}
\int f(y) K^2(x-y)\, dy &\leq \int_{y < x/2} f(y) K^2\left(\frac{x}{2}\right) dy \\
&\quad + \int_{x/2 \leq y} f\left(\frac{x}{2}\right) K^2(x-y)\, dy \\
&\leq K^2\left(\frac{x}{2}\right) + f\left(\frac{x}{2}\right) \int K^2.
\end{aligned}$$

the square root of the right-hand side does not exceed $|K(x/2)| + \sqrt{f(x/2)} \sqrt{\int K^2}$. Integration with respect to x gives the stated inequality.

THEOREM 13. *Let $s \geq 1$ be an integer, and let K be a symmetric bounded function with compact support, satisfying the conditions of Lemma 22. Let L*

be as in Lemma 22, and let C_s be the constant of Lemma 25. For densities f satisfying condition C of Lemma 26, the kernel estimate f_n satisfies the following inequality as $h \downarrow 0$:

$$E\left(\int |f_n - f|\right) \le C_s \int |L| h^s + (nh)^{-1/2} \int \sqrt{f} \sqrt{\int K^2} (1 + o(1)).$$

In particular, if $C_s < \infty$ and $\int \sqrt{f} < \infty$ (i.e., $D_s^*(f) < \infty$), and

$$h = \left(\frac{\sqrt{\int K^2} \int \sqrt{f}}{2sC_s \int |L|} \right)^{2/(2s+1)} n^{-1/(2s+1)},$$

then

$$E\left(\int |f_n - f|\right) \le \frac{2s+1}{2s} \left(2s \int |L| \left(\sqrt{\int K^2} \right)^{2s} n^{-s} \right)^{1/(2s+1)}$$

$$\times D_s^*(f)(1 + o(1)).$$

Proof. Theorem 13 follows without work from Theorem 11 and Lemma 26, part C.

Thus, even for f with unbounded support, $D_s^*(f)$ seems to appear as the measure of difficulty. The most important cases are again $s = 1$ and $s = 2$. In those situations, the kernel K that minimizes the bound is independent of f, and coincides once again with the isosceles triangular density for $s = 1$ and the Epanechnikov kernel for $s = 2$.

With those choices for K, the upper bounds become

$$\left(\frac{3}{2n}\right)^{1/3} D_1^*(f) = \left(\frac{3}{n}\right)^{1/3} B_H^*(f) \qquad (s = 1)$$

and

$$\left(\tfrac{225}{512}\right)^{1/5} n^{-2/5} D_2^*(f) = \left(\tfrac{225}{256}\right)^{1/5} n^{-2/5} B^*(f) \qquad (s = 2).$$

These are only fractionally larger than the corresponding upper bounds of Theorems 10 and 1. The optimal values for h differ also very little from those obtained for compact support densities in Theorems 10 and 1.

We finish this section by noting that condition C of Lemma 26 holds for all bounded unimodal almost everywhere continuous f with $\int\sqrt{f} < \infty$: indeed, if m is a mode, then for $x > m$,

$$\sup_{|y-x|\le t} f(y) \le f(\max(x-t, m)),$$

and the square root of this is integrable for any t.

11. UNBIASEDNESS AND THE ACHIEVABILITY OF THE ERROR RATE $1/\sqrt{n}$

The kernel estimate has an expected L_1 error rate that decreases as $n^{-s/(2s+1)}$ for all f in $F_{s,r}$ or $W(s-1,1,C)$ (see Theorem 11), provided that K and h are picked appropriately. By increasing s, these classes become smaller, and the rate $n^{-1/2}$ is approached. We also recall that $1/\sqrt{n}$ is the best possible minimax rate of convergence for any density estimate over such simple one-parameter classes as $Q_1(g)$ (Theorem 4.8) or $\Pi(g)$ (Theorem 4.4). In between, there is a void: for some estimate f_n, does there exist a rich class of densities $\mathcal{F}$ for which $\limsup_{n\to\infty}\sqrt{n}\,E(\int|f_n - f|) < \infty$, all $f \in \mathcal{F}$? By "rich," we mean that the class should certainly not be describably by a finite number of parameters, although, from the lower bounds of Chapter 4, it should be clear that $\mathcal{F}$ cannot be too large.

The answer is affirmative in L_2. In fact, Ibragimov and Khasminskii (1982) have obtained the following fantastic theorem:

THEOREM 14 (Ibragimov and Khasminskii, 1982). *Let A_T be the class of all densities with characteristic function vanishing outside $[-T, T]$. Then, if f_n denotes a density estimate,*

$$\liminf_{n\to\infty}\inf_{f_n}\sup_{f\in A_T} nE\left(\int(f_n - f)^2\right) = \frac{T}{\pi}.$$

A few comments are in order here. First, the class A_T consists of extremely smooth densities because tail conditions on characteristic functions correspond to smoothness conditions on the density. Since the characteristic function ϕ is absolutely integrable, we can obtain f from ϕ by inversion (see, e.g., Lukacs, 1970):

$$f(x) = (2\pi)^{-1}\int e^{-itx}\phi(t)\,dt,$$

and in fact, we can obtain $f^{(s)}$ by taking the sth derivative inside the integral. This gives

$$\int |f^{(s)}| \leq \frac{1}{\pi} \frac{T^{s+1}}{s+1}, \quad \text{all integer } s.$$

Unfortunately, there exist no direct inequalities between the L_1 and the L_2 errors, so that Theorem 14 does not imply

$$\limsup_{n \to \infty} \sqrt{n} \sup_{f \in A_T} E\left(\int |f_n - f| \right) < \infty$$

for some estimate f_n. There are a few indirect inequalities but they are of no help. For example, we have inequalities of the form given in Theorem 8.3.

The possibility of achieving the rate $n^{-1/2}$ within A_T should come as no surprise because the class is "nearly" parametric: by Nyquist's theorem, we known that f can be completely reconstructed (Feller, 1971) from the value of f at the points iy, $i = 0, +1, -1, +2, -2, \ldots$, where y is a small enough positive constant. A_T can be considered therefore as a class with a countable number of parameters.

Ibragimov and Khasminskii have shown more: the bound of Theorem 14 is achieved for the *Fourier integral estimate* (FIE) described in Davis (1975, 1977) and Konakov (1973). The achievability of the $1/n$ error rate in L_2 for $f \in A_T$ on an individual basis was also noted by Davis (1975, 1977), and is based upon the L_2 analysis of Watson and Leadbetter (1963). The FIE is a kernel estimate with kernel

$$K(x) = \frac{\sin x}{\pi x}.$$

Note that $\int K = 1$, but that $\int |K| = \infty$. Also, K has characteristic function $I_{[-1,1]}(t)$. The estimate f_n thus obtained as the correct integral (1), but it is not absolutely integrable with probability 1. Hence, we cannot "normalize" f_n by defining

$$f_n^*(x) = \frac{f_n(x) I_{[f_n(x) > 0]}}{\int_{f_n > 0} f_n},$$

because f_n^* would be zero with probability 1. Had f_n been absolutely integrable, then this normalization would have led to a valid density f_n^* and

moreover,

$$\int |f_n^* - f| \le \int |f_n - f|.$$

(Apply the nonnegative projection Theorem 11.4). Since we do not know of any other normalizations that keep or reduce the L_1 error, we are thus reluctant to recommend the FIE as a density estimate, in keeping with the general principles established at the outset of this book. The reason why Davis and Ibragimov and Khasminskii were able to obtain a $1/n$ rate in L_2 was that, for each x, the entire sample helped in the estimation. For the kernel estimate, this would imply that h does not tend to 0. But this in turn would imply that $\int |f - f^* K_h| = 0$ for some positive h, that is, the estimate is unbiased! If we try to follow this reasoning, then the key to the solution is the existence of a function K with the property that $\int K = 1$, $\int |K| < \infty$ (for normalization), and $\int |f - f^* K_h| = 0$ for some $h > 0$ and all $f \in A_T$. Such a K indeed exists, so that there is hope to obtain a $1/\sqrt{n}$ expected L_1 error rate on A_T.

We start with the *de la Vallée Poussin density*

$$K(x) = (2\pi)^{-1}\left(\frac{\sin(x/2)}{x/2}\right)^2,$$

which has characteristic function $(1 - |t|)_+$. Then define for a constant $a > 0$ (to be picked later) the function g_a:

$$g_a(x) = (a+1)\left(K(x) - K\left(\frac{a+1}{a}x\right)\right).$$

We see that $\int g_a = 1$, and that $\int |g_a| \le 2a + 1$. Also, g_a has characteristic function

$$\psi_a(t) = (a+1)(1 - |t|)_+ - a\left(1 - \left|\frac{a+1}{a}t\right|\right)_+,$$

which is

$$\psi_a(t) = \begin{cases} 1, & |t| \le a/(1+a), \\ (a+1)(1-|t|), & a/(1+a) \le |t| \le 1 \\ 0, & |t| \ge 1. \end{cases}$$

We can now handle the unbiasedness of our kernel estimate on A_T:

THEOREM 15. *Let A_T be the class of all densities with characteristic function vanishing outside $[-T, T]$. We have an unbiased kernel estimate with kernel K and smoothing factor h (i.e., $\int |f - f^* K_h| = 0$) in any one of the following cases:*

1. $K(x) = (\sin x)/\pi x$, $h \le 1/T$.
2. $K(x) = g_a(x)$ *for fixed* $a > 0$, $h \le (a/(1+a))(1/T)$.

Proof. Let f have characteristic function ϕ. It is known that $f * K_h$ has characteristic function $\phi(t)\psi(th)$, where ψ is the characteristic function of K (which, when K is not a density, is defined as $\int e^{itx} K(x)\,dx$). For $f = f * K_h$ for almost all x, it suffices that $\phi(t) = \phi(t)\psi(th)$ for all t. Since $f \in A_T$, we need only verify that $\psi(th) = 1$, all $|t| \le T$. For the first kernel, we have $\psi(t) = I_{[1,1]}(t)$, and we need only require that $h \le 1/T$ (this argument is due to Davis (1975, 1977)). For the second kernel, we have to ask that $Th \le a/(1+a)$. This concludes the proof of Theorem 15.

Before we proceed with the properties of the kernel estimate with kernel g_a for $f \in A_T$, we will see what we should not expect, and what we cannot do:

THEOREM 16. *Let f_n be a kernel estimate with kernel K satisfying $\int K = 1$, $\int |K| < \infty$, and let $\int \sup_{|u| \ge |x|} |K(u)| < \infty$. Then,*

1. $h \to 0$ *implies* $\sqrt{n}\, E(\int |f_n - f|) \to \infty$, *all* f.
2. $h \to 0$, $nh \to \infty$, *and* $\int \sqrt{f} = \infty$ *imply* $\sqrt{nh}\, E(\int |f_n - f|) \to \infty$.
3. *If the characteristic function ϕ of f is nonzero except possibly on a set of Lebesgue measure 0, then* $\sqrt{n} \inf_{h>0} E(\int |f_n - f|) \to \infty$.
4. *If K is a density (but possibly without integrable radial majorant), then* $\inf_f \liminf_{n\to\infty} n^{2/5} \inf_{h>0} E(\int |f_n - f|) > 0$.

Theorem 16 states that if we are to construct a kernel estimate that is "normalizable" and consistent for *all* f, then we must have $h \to 0$, and *in that case, the* $1/\sqrt{n}$ *rate is not achievable even for a single* f! Properties 1 and 3 of Theorem 16 essentially imply that it is useless to look outside A_T for $1/\sqrt{n}$ error rates. Finally, property 4, which coincides with Theorem 2, gives an even stronger lower bound for individual densities f when we do not allow negative-valued kernels. The proof of Theorem 16 rests on the following uniform lower bounds:

LEMMA 27. *Let $Z_1, \ldots, Z_n$ be independent identically distributed zero mean random variables with $E(|Z_1|^p) < \infty$, for some $p \ge 2$. Then there are positive constants B_p and C_p only depending upon p such that*

$$B_p E\left(\left(\sum_{i=1}^n Z_i^2\right)^{p/2}\right) \le E\left(\left|\sum_{i=1}^n Z_i\right|^p\right) \le C_p E\left(\left(\sum_{i=1}^n Z_i^2\right)^{p/2}\right).$$

Furthermore,

$$E\left(\sqrt{\sum_{i=1}^{n} Z_i^2}\right) \geq E\left(\left|\sum_{i=1}^{n} Z_i\right|\right) \geq \sqrt{\frac{n}{8}}\, E(|Z_1|).$$

We also have, putting $q = P(|Z_1| \geq u)$, $u > 0$:

$$E\left(\left|\sum_{i=1}^{n} Z_i\right|\right) \geq u\frac{\sqrt{nq}}{4\sqrt{2}}\,\frac{nq}{8+nq}$$

$$\geq \frac{u\sqrt{nq}}{20\sqrt{2}}, \quad \text{valid if } nq \geq 2.$$

REMARK. The first inequality of Lemma 27 is due to Marcinkiewicz and Zygmund (1937) (see, e.g., Manstavicius, 1982, and the references found there). We will only prove the other inequalities of Lemma 27.

Proof of Lemma 27. We will repeatedly use the following inequality: if U, V are arbitrary random variables, then $E(|U+V| \geq E(|U+E(V|U)|)$. We first symmetrize our problem by using the fact that if U, V are independent identically distributed random variables, then

$$E(|U|) = \tfrac{1}{2}E(|U|+|V|) \geq \tfrac{1}{2}E(|U-V|).$$

Thus,

$$E\left(\left|\sum_{i=1}^{n} Z_i\right|\right) \geq \frac{1}{2}E\left(\left|\sum_{i=1}^{n}(Z_i - Z_i')\right|\right),$$

where $Z_1', \ldots, Z_n', Z_1, \ldots, Z_n$ are independent and identically distributed. By Szarek's inequality (Lemma 28) and Jensen's inequality, the lower bound is at least equal to

$$\frac{1}{2}\frac{1}{\sqrt{2}}E\left(\left|\sum_{i=1}^{n}(Z_i - Z_i')^2\right|\right) \geq \sqrt{\frac{n}{8}}\, E(|Z_1 - Z_1'|) \geq \sqrt{\frac{n}{8}}\, E(|Z_1|).$$

This can be seen by representing $Z_i - Z_i'$ as $R_i B_i$, where $R_i = |Z_i - Z_i'|$ and $B_i = \operatorname{sign}(Z_i - Z_i')$ are independent, and by conditioning on the R_i's.

The last inequality of Lemma 27 can be obtained as follows. Set $v = E(R_1|R_1 \geq u)$, note that $v \geq u$, and define $N = \sum_{i=1}^{n} I_{[R_i \geq u]}$. We have

$$E\left(\left|\sum_{i=1}^{n} R_i B_i\right|\right) = E\left(\left|\sum_{i=1}^{n} B_i\left(R_i I_{[R_i \geq u]} + R_i I_{[R_i < u]}\right)\right|\right)$$

$$\geq E\left(\left|\sum_{i=1}^{n} B_i\left(v I_{[R_i \geq u]} + E(R_1|R_1 < u) I_{[R_i < u]}\right)\right|\right)$$

(this is obtained by conditioning on $B_i, I_{[R_i \geq u]}$ and applying the conditional form of Jensen's inequality)

$$= E\left(\left|\sum_{i=1}^{N} v B_i + \sum_{j=1}^{n-N} B_{j+N} E(R_1|R_1 < u)\right|\right)$$

$$\geq E\left(\left|\sum_{i=1}^{N} v B_i\right|\right) \quad \text{(by independence, given } N\text{)}$$

$$\geq uE\left(\left|\sum_{i=1}^{N} B_i\right|\right)$$

$$\geq uE\left(\sqrt{N/2}\right) \quad \text{(by Szarek's inequality).}$$

Let us define $r = P(|Z_1 - Z_1'| \geq u)$ and note that $r \geq \frac{1}{2}P(|Z_1| \geq u) = q/2$. Because N is binomial (n, r), we have by Cantelli's form of Chebyshev's inequality for $nr \geq 1$:

$$E\left(\left|\sum_{i=1}^{n} Z_i\right|\right) \geq uE\left(\sqrt{N/8}\right) \geq u\sqrt{nr}\, P(N \geq nr/2) \cdot \frac{1}{4}$$

$$\geq u\frac{\sqrt{nr}}{4} \frac{(nr/2)^2}{(nr/2)^2 + nr(1-r)} \geq u\frac{\sqrt{nr}}{4} \frac{nr}{4 + nr} \geq u\frac{\sqrt{nr}}{20}$$

$$\geq u\frac{\sqrt{nq}}{20\sqrt{2}},$$

which was to be shown.

LEMMA 28 (Inequalities for the Binomial Distribution). (*Khinchine's inequality*). *Let $Y_1, \ldots, Y_n$ be independent random variables taking the values*

$+1$ and -1 with equal probability, and let $a_1, \ldots, a_n$ be real numbers. Then there exist positive constants B_p and C_p depending upon $p > 0$ only such that

$$B_p \sqrt{\sum_{i=1}^n a_i^2} \le E^{1/p}\left(\left|\sum_{i=1}^n a_i Y_i\right|^p\right) \le C_p \sqrt{\sum_{i=1}^n a_i^2}.$$

The following values are optimal:

$$B_p = 2^{1/2-1/p}, \qquad 0 < p \le p_0.$$

$$B_p = 2^{1/2}\left(\frac{\Gamma((p+1)/2)}{\sqrt{\pi}}\right)^{1/p}, \qquad p_0 \le p \le 2.$$

$$B_p = 1, \qquad p \ge 2.$$

$$C_p = 2^{1/2}\left(\frac{\Gamma((p+1)/2)}{\sqrt{\pi}}\right)^{1/p}, \qquad 2 \le p.$$

$$C_p = 1, \qquad 0 < p \le 2.$$

Here $p_0 = 1.84742\cdots$ is the solution in $(1,2)$ of $\Gamma(p+1)/2) = \Gamma(\frac{3}{2})$.

REMARK. The optimal constant $B_1 = 1/\sqrt{2}$ was obtained by Szarek (1976) (and we will refer to the corresponding inequality as Szarek's inequality). The best values for C_p, $p \ge 3$, are due to Young (1976). All the cases are treated simultaneously in Haagerup (1978). We note in passing that if Y is binomial $(n, \frac{1}{2})$, then $E(|Y - n/2|) \ge \sqrt{n/8}$.

LEMMA 29. *Let f be a density with characteristic function ϕ, and let K be a Borel measurable function satisfying $\int K = 1$, $\int |K| < \infty$, and assume that K has characteristic function $\psi(t) = \int e^{itx} K(x)\,dx$, $t \in R$. Then,*

$$\int |f - f^* K| \ge \sup_t |\phi(t) - \phi(t)\psi(t)|.$$

If $c > 0$ is a constant and $h \to c$, then $\int |f * K_h - f * K_c| \to 0$.

Proof.

$$\sup_t |\phi - \phi\psi| = \sup_t \left|\int (f(x) - (f * K)(x)) e^{itx}\,dx\right|$$

$$\le \int |f(x) - (f * K)(x)| \sup_t |e^{itx}|\,dx = \int |f - f * K|.$$

For the second statement of Lemma 29, verify that (2.4) in the proof of Theorem 2.4 remains valid.

Proof of Theorem 16. Statement 4 is contained in Theorem 2. For the other statements we shall use the crude inequalities (see, e.g., Theorem 3.6):

$$E\left(\int |f_n - f|\right) \geq \int |f - f^* K_h|, \tag{21}$$

$$E\left(\int |f_n - f|\right) \geq \frac{1}{2} E\left(\int |f_n - f * K_h|\right). \tag{22}$$

Assume that we have shown 1. Then 3 follows by contradiction. Indeed, if there exists a sequence h such that $\sqrt{n}\, E(\int |f_n - f|)$ remains bounded, then there exists a constant $c > 0$ for which $h \to c$ along this subsequence. By Lemma 29, we conclude that $\sup_t |\phi(t) - \phi(t)\psi(ct)| = 0$. But this implies that $\psi(t) = 1$, that is, the measure corresponding to ψ is atomic with mass 1 at the origin, and this is our contradiction.

Consider now 1 and 2 together. Lemma 27 will be applied with $Z_i = Z_i(x) = (K_h(x - X_i) - E(K_h(x - X_i)))/n$. We note that $E(Z_i(x)) = 0$, and that

$$E\left(\int |f_n - f * K_h|\right) = \int E\left(\left|\sum_{i=1}^{n} Z_i(x)\right|\right) dx$$

$$\geq \int_B \frac{a}{nh} \sqrt{n}\, (20\sqrt{2})^{-1} \sqrt{P\left(|Z_1(x)| \geq \frac{a}{nh}\right)}\, dx,$$

where $a > 0$ is a number to be chosen further on, and B is the set of all x for which $P(|Z_1(x)| \geq a/nh)$ is at least equal to $2/n$. Here we used Lemma 27. Thus,

$$\sqrt{nh}\, E\left(\int |f_n - f|\right) \geq \frac{a}{40\sqrt{2h}} \int_B \sqrt{P\left(|Z_1(x)| \geq \frac{a}{nh}\right)}\, dx.$$

Let C be the set of all x for which $|K(x)| \geq a/2$, and let $a > 0$ be so small that $\int_C dx = b > 0$ for some positive b. By Theorem 2.3, $E(|K_h(x - X_1)|) = f * |K|_h \to f\int |K|$ for almost all x. For such x, and for all h small

enough (so that at least $E(|K_h(x - X_1)|) \le a/2h$), we have

$$P\left(|Z_1(x)| \ge \frac{a}{nh}\right) \ge P\left(|K_h(x - X_1)| \ge \frac{a}{h} - E(|K_h(x - X_1)|)\right)$$

$$\ge P\left(|K_h(x - X_1)| \ge \frac{a}{2h}\right) = P\left(|K\left(\frac{x - X_1}{h}\right)| \ge \frac{a}{2}\right)$$

$$= P(X_1 \in x - hC)$$

$$\sim f(x)h\int_C dx \quad \text{(almost all } x\text{; Theorem 2.3)}$$

$$= f(x)hb.$$

Thus, by Fatou's lemma, if $nh \to \infty$,

$$\liminf_{n\to\infty} \sqrt{nh}\, E\left(\int |f_n - f|\right) \ge \frac{a}{40\sqrt{2}} \int \liminf_{n\to\infty} I_B \sqrt{\frac{P(|Z_1(x)| > a/nh)}{h}}\, dx$$

$$\ge \frac{a}{40\sqrt{2}} \int_{f>0} b\sqrt{f(x)}\, dx.$$

From this, we immediately deduce statement 2, and part of statement 1.

This leaves us with the case $\liminf_{n\to\infty} nh < \infty$ in statement 1. Obviously, we can assume that $\limsup_{n\to\infty} nh \le C_0 < \infty$, for the case $\limsup_{n\to\infty} nh = \infty$ can then be obtained by a trivial combination of arguments. By the last inequality of Lemma 27,

$$\sqrt{n}\, E\left(\int |f_n - f|\right) \ge \frac{a}{4\sqrt{2}} \int \frac{1}{h}\sqrt{q}\, \frac{nq}{8 + nq}\, dx,$$

where $q = P(|Z_1(x)| \ge a/nh) \ge f(x)hb(1 + o(1))$ for almost all x. By Fatou's lemma, the limit infimum of the left-hand side is ∞ if, for all constants $c > 0$,

$$\liminf_{n\to\infty} \frac{nh}{c + nh} \frac{1}{\sqrt{h}} = \infty.$$

But this follows from $\liminf nh > 0$, $h \to 0$. This leaves us with the case $h = o(1/n)$. We will conclude the proof of statement 1 and Theorem 16 if

we can show that $h = o(1/n)$ implies

$$\liminf_{n\to\infty} E\left(\int |f - f_n|\right) \geq 1.$$

(Note that this does not follow from Theorem 3.1 since we allow negative-valued kernels.) Now,

$$E\left(\int |f_n - f|\right) \geq E\left(\int (f - f_n)_+\right) = E\left(\int \left(f I_{[f_n \leq f]} - f_n I_{[f_n \leq f]}\right)\right)$$

$$= \int f P(f_n \leq f) - \int E\left(f_n I_{[f_n \leq f]}\right).$$

By the Lebesgue dominated convergence theorem, we are done if we can show that for almost all x (with respect to f), $f_n \to 0$ in probability. Now, to do so, we define $C = \{x: |K(x)| \geq \varepsilon\}$, where $\varphi > 0$ is arbitrary, and note that $|f_n| \leq \varepsilon$ if $x + hC$ has no X_i's. But

$$P(x + hC \text{ contains at least one } X_i) \leq n \int_{x+hC} f \to nhf(x) \int_C dy$$

for almost all x. (This would follow from Theorem 2.3 if the function I_C had an integrable radial majorant, and this is of course a simple consequence of the fact that K has an integrable radial majorant.) Now, $\int_C dx \leq \int |K|/\varepsilon < \infty$, and thus, we are done, because ε was arbitrary and $nh = o(1)$.

Because the factor $\int \sqrt{f}$ is infinite for many a density in A_T (such as the de la Vallée Poussin density; in fact, one can show that it is infinite for all densities with real even characteristic functions that are concave on $[0, T]$ and vanish outside $[-T, T]$) for us to be able to use the uniform bound of Lemma 25 and inequality D of Lemma 26, we must require explicitly that $\int \sqrt{f}$ be finite. For this reason, we introduce the slightly smaller class of densities $A_{T,s,C}$, where T, C are positive constants and $s \geq 1$ is an integer:

$$A_{T,s,C} = \Big\{ f: f \text{ has characteristic function } \phi, \text{ where}$$
$$\text{(i) } \phi = 0 \text{ outside } [-T, T] \text{ (i.e., } f \in A_T);$$
$$\text{(ii) } \phi, \ldots, \phi^{(s-1)} \text{ exist and are absolutely continuous (with almost everywhere derivatives } \phi^{(1)}, \ldots, \phi^{(s)});$$
$$\text{(iii) } \int |\phi^{(s)}| \leq C. \Big\}$$

This class is not empty and certainly not describable by a finite number of parameters. For example, it contains all densities whose characteristic functions are obtained by convolving $(1 - |t|)_+$ sufficiently often with itself (this will approach a normal characteristic function), and normalizing (so that the value at 0 is 1). Ignoring scale factors, the densities will be of the form

$$C_p\left(\frac{\sin x}{x}\right)^{2p}$$

for integer p. Of course, it is clear that for some values of C and T, $A_{T,s,C}$ is indeed empty. We should stress from the beginning that we are not interested in $A_{T,s,C}$ per se since it seems a rather artificial and unrealistic class, but in the mere existence of a sufficiently rich class of densities $\mathcal{F}$ for which we can construct an estimate satisfying

$$\limsup_{n\to\infty} \sup_{f\in\mathcal{F}} \sqrt{n}\, E\left(\int |f_n - f|\right) < \infty.$$

Thus, consider the kernel estimate with kernel $K(x) = g_a(x)$, and let us call it the *trapezoidal kernel estimate* because g_a has a trapezoidal characteristic function. Let $a > 0$ be fixed, and let $h > 0$ be fixed such that

$$h \le \frac{a}{1+a}\frac{1}{T}. \tag{23}$$

Then, by Lemma 25 and Theorem 15,

$$E\left(\int |f_n - f|\right) = E\left(\int |f_n - f * K_h|\right) \le (nh)^{-1/2} \int \sqrt{f * (K^2)_h}. \tag{24}$$

The last integral in (24) will now be bounded simply by obtaining a uniform upper bound for all f in $A_{T,s,C}$:

LEMMA 30 (Inequalities Linking f and Its Characteristic Function ϕ). *For* all *f with characteristic function ϕ,*

$$\sup f \le \frac{1}{2\pi}\int |\phi|.$$

For all $f \in A_{T,s,C}$,

$$|x|^s f \le \frac{1}{2\pi}\int |\phi^{(s)}| \le \frac{C}{2\pi},$$

and

$$f(x) \le g(x) \triangleq \min\left(\frac{T}{\pi}, \frac{C}{2\pi|x|^s}\right).$$

Proof. The first inequality follows from

$$f(x) = (2\pi)^{-1}\int e^{-itx}\phi(t)\,dt \le (2\pi)^{-1}\int|\phi|.$$

Note that for $f \in A_T$, this does not exceed T/π. Next, by partial integration of the inversion formula, and the absolute continuity of $\phi, \ldots, \phi^{(s-1)}$,

$$f(x) = (2\pi)^{-1}\int (ix)^{-s}e^{-itx}\phi^{(s)}(t)\,dt$$

$$\le (2\pi|x|^s)^{-1}\int|\phi^{(s)}|.$$

This concludes the proof of Lemma 30.

The class $A_{T,s,C}$ is not closed under translations, for otherwise there would not exist an integrable uniform bound for f. To define a translation invariant class is easy of course, but we will not go through the extra trouble here. Note that $\int|\phi|$ tells us something about the peak of f, and that $\int|\phi^{(s)}|$ gives us information about a uniform upper bound for the tail of f. We can now state the last result of this section:

THEOREM 17. *Let $s \ge 3$ be integer, and let T and C be large enough so that $A_{T,s,C}$ is nonempty. Then, the trapezoidal kernel estimate f_n with smoothing factor h chosen fixed as in* (23) *and $k = g_1$ satisfies*:

$$\sup_{f\in A_{T,s,C}} \sqrt{n}\,E\left(\int|f_n - f|\right) \le \left(\frac{16s}{s-2}\frac{T^{1/2-1/s}}{\pi^{3/2}\sqrt{2}}\left(\frac{C}{2}\right)^{1/s} + \frac{4}{\sqrt{\pi}}\right)\frac{1}{\sqrt{h}}$$

Proof. For positive α, β, γ, we verify first the integral

$$\int\min\left(\alpha, \frac{\beta}{|x|^\gamma}\right) = \frac{2\gamma}{\gamma-1}\alpha^{1-1/\gamma}\beta^{1/\gamma}.$$

Now, apply the following inequalities to (24): $f \le g$ as defined in Lemma 30, and $|K(x)| \le 2\min(1/2\pi, (4/x^2)(1/\pi)) = \min(1/\pi, (4/\pi)x^{-2}) =$

$K^*(x)$ (by definition of K^*). Also, $f*(K^2)_h \le g*(K^{*2})_h$. Since both g and K^* are symmetric and unimodal, we can apply Lemma 26, part D. The result follows after observing that

$$\int |K^*| = \frac{8}{\pi}, \qquad \int K^{*2} = \frac{16}{3\pi^2},$$

$$\int \sqrt{g} = \frac{2s}{s-2} T^{1/2-1/s} \left(\frac{C}{2}\right)^{1/s} \pi^{-1/2}.$$

To make the inequality as small as possible, it is necessary to choose h as large as possible. With the value $h = 1/2T$ on the right-hand side of the inequality, we obtain a minimax upper bound for $A_{T,s,C}$. For every value of a in the trapezoidal kernel estimate, we obtain a different minimax bound. The formal optimization of it with respect to a is not given here.

What we retain from this section is that no kernel estimate with deterministic h can be consistent for all densities unless $h \to 0$. Thus, we have not constructed a density estimate that is consistent for all f and satisfies

$$\limsup_{n\to\infty} \sqrt{n}\, E\left(\int |f_n - f|\right) \le c, \quad \text{all } f \in \mathscr{F},$$

where $\mathscr{F}$ is a rich enough family of densities. Another unanswered question is whether

$$\limsup_{n\to\infty} \sup_{f\in A_T} \sqrt{n}\, E\left(\int |f_n - f|\right) < \infty$$

for some estimate f_n.

The trapezoidal kernel estimate has a kernel with $\int x^i K = 0$ for all $i > 0$. The intriguing property of this estimate is that we do not have to adjust K to the smoothness of f as in Bartlett's estimates. Of course, h still needs adjusting according to the smoothness of f, or alternatively, h can be estimated from the data by one of the methods described in Chapter 6.

Finally, if we follow the interesting L_2 theory developed by Watson and Leadbetter (1963) and picked up again by Davis (1975, 1977), we can consider what rates of convergence are attainable under various conditions on the tail of ϕ. In doing so, we can obtain a continuum of rates between $1/\sqrt{n}$ and $n^{-2/5}$ (such as $\log n/\sqrt{n}$, etc.) depending upon the rate of decrease of the tail of $|\phi|$. An enumeration of the standard tail conditions

would be tantamount to admitting that these lead to important classes: they do not, and $A_{T,s,C}$ is not important either. Rare are the occasions when we know anything about the smoothness of f, and expensive is the price paid for choosing the "wrong" h in the kernel estimate! (For example, if we "gamble" that f is in $A_{T,s,C}$, we can choose a fixed h and employ the trapezoidal kernel estimate, and if we are wrong, the punishment is severe: f_n may not converge at all to f.)

REFERENCES

S. Abou-Jaoude (1977). La convergence L_1 et L_∞ de certains estimateurs d'une densité de probabilité, Thèse de Doctorat d'État, Université de Paris VI, Paris, France.

M. S. Bartlett (1963). Statistical estimation of density functions, *Sankhya Series A* **25**, pp. 245–254.

E. F. Beckenbach and R. Bellman (1965). *Inequalities*, Springer-Verlag, Berlin.

J. Bretagnolle and C. Huber (1979). Estimation des densités: risque minimax, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **47**, pp. 119–137.

F. Carlson (1934). Une inégalité, *Arkiv foer Matematik, Astromi och Fysik* **25B**, pp. 1–15.

K. B. Davis (1975). Mean square error properties of density estimates, *Annals of Statistics* **3**, pp. 1025–1030.

K. B. Davis (1977). Mean integrated square error properties of density estimates, *Annals of Statistics* **5**, pp. 530–535.

P. Deheuvels (1977). Estimation non paramétrique de la densité par histogrammes généralisés, *Revue de Statistique Appliquée* **25**, pp. 5–42.

P. Deheuvels and P. Hominal (1980). Estimation automatique de la densité, *Revue de Statistique Appliquée* **28**, pp. 25–55.

L. Devroye and C. S. Penrod (1982). Distribution-free lower bounds in density estimation, Technical Report, Applied Research Laboratories, The University of Texas at Austin.

V. A. Epanechnikov (1969). Nonparametric estimates of a multivariate probability density, *Theory of Probability and Its Applications* **14**, pp. 153–158.

W. Feller (1968). *An Introduction To Probability Theory and Its Applications*, Vol. 1, Wiley, New York.

W. Feller (1971). *An Introduction to Probability Theory and Its Applications*, Vol. 2, Wiley, New York.

D. Freedman and P. Diaconis (1981). On the histogram as a density estimator: L_2 theory, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **57**, pp. 453–476.

U. Haagerup (1978). Les meilleures constantes de l'inégalite de Khintchine, *Comptes Rendus de l'Académie des Sciences de Paris A* **286**, pp. 259–262.

I. A. Ibragimov and R. Z. Khasminskii (1982). Estimation of distribution density belonging to a class of entire functions, *Theory of Probability and Its Applications* **27**, pp. 551–562.

V. D. Konakov (1973). Nonparametric estimation of density functions, *Theory of Probability and Its Applications* **17**, pp. 361–362.

G. Kulldorf (1963, 1964). On the optimum spacing of sample quantiles from a normal distribution, Part 1, *Skandinavisk Aktuarietidskrift* **46**, pp. 143–161, 1963; Part 2, *Skandinavisk Aktuarietidskrift* **47**, pp. 71–87, 1964.

E. Lukacs (1970). *Characteristic Functions*, Griffin, London.

E. Manstavicius (1981). Inequalities for the pth moment, $0 < p < 2$, of a sum of independent random variables, *Lithuanian Mathematical Journal* **22**, pp. 64–67.

J. Marcinkiewicz and A. Zygmund (1937). Sur les fonctions indépendantes, *Fundamentales de Mathematiques* **29**, pp. 60–90.

E. Parzen (1962). On estimation of a probability density function and the mode, *Annals of Mathematical Statistics* **33**, pp. 1065–1076.

V. V. Petrov (1975). *Sums of Independent Random Variables*, Springer-Verlag, Berlin.

M. Rosenblatt (1956). Remarks on some nonparametric estimates of a density function, *Annals of Mathematical Statistics* **27**, pp. 832–835.

M. Rosenblatt (1971). Curve estimates, *Annals of Mathematic Statistics* **42**, pp. 1815–1842.

M. Rosenblatt (1979). Global measures of deviation for kernel and nearest neighbor density estimates, in *Smoothing Techniques for Curve estimation*, Th. Gasser and M. Rosenblatt (Eds.), Lecture Notes in Mathematics #757, pp. 181–190, Springer-Verlag, Berlin.

D. W. Scott (1979). On optimal data-based histograms, *Biometrika* **66**, pp. 605–610.

S. J. Szarek (1976). On the best constants in the Khintchine inequality, *Studia Mathematica* **63**, pp. 197–208.

R. A. Tapia and J. R. Thompson (1978). *Nonparametric Probability Density Estimation*, The Johns Hopkins University Press, Baltimore.

G. S. Watson and M. R. Leadbetter (1963). On the estimation of the probability density, *Annals of Mathematical Statistics* **34**, pp. 480–491.

W. Wertz (1972). Fehlerabschätzung für eine Klasse von nichtparametrischen Schätzfolgen, *Metrika* **19**, pp. 131–139.

E. T. Whittaker and G. N. Watson (1963). *A Course of Modern Analysis*, 4th ed., Cambridge University Press, Cambridge, U.K.

R. M. G. Young (1976). On the best possible constants in the Khintchine inequality, *Journal of the London Mathematical Society* **14**, pp. 496–504.

CHAPTER 6

The Automatic Kernel Estimate: L_1 and Pointwise Convergence

1. THE MAIN RESULT

In this chapter, we study the consistency of the *automatic kernel estimate*

$$f_n(x) = (nh^d)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{1}$$

where K is a fixed density and $h = h(n, X_1, \ldots, X_n)$ is a Borel measurable function of n and the data. Ideally, h does not depend upon parameters that have to be chosen by the user, although, strictly speaking, such estimates (including the standard kernel estimate) are called automatic kernel estimates as well. Note that h is not allowed to depend upon x since this would in general lead to an estimate that is no longer a density on R^d.

The first and foremost result of this chapter is in the spirit of Theorem 3.1 for the standard kernel estimate. It is stated here without proof (see Section 5 for the proof). In Section 4 we will present several examples of automatic kernel estimates. In Sections 2 and 3, the pointwise convergence properties of these estimates are studied.

THEOREM 1. *Let f_n be an automatic kernel estimate with arbitrary density K. If $h + (nh^d)^{-1} \to 0$ completely (almost surely, in probability) then $\int |f_n - f| \to 0$ completely (almost surely, in probability), for all densities f on R^d.*

2. POINTWISE CONVERGENCE OF THE AUTOMATIC KERNEL ESTIMATE

Another albeit less powerful way of proving the L_1 consistency of a density estimate consists of establishing the pointwise convergence of the estimate

at almost all x, and applying Scheffé's Theorem 2.7 or Glick's extension of it (Theorem 2.8) to derive the L_1 consistency. For the pointwise convergence of the automatic kernel estimate and all densities f on R^d we have the following theorem:

THEOREM 2. *Let K be a Riemann integrable density with compact support, and let f_n be an automatic kernel estimate with smoothing factor h. Let f be a fixed but arbitrary density on R^d. Then*:

A. If $h + (nh^d)^{-1} \to 0$ in probability, then $f_n \to f$ in probability at almost all x.

B. If $h \to 0$ and $nh^d/(\log\log n) \to \infty$ almost surely, then $f_n \to f$ almost surely at almost all x.

C. If $h \to 0$ and $nh^d/(\log n) \to \infty$ completely, then $f_n \to f$ completely at almost all x.

Theorem 2 is proved in Section 5. A and B lead directly to L_1 consistency in probability and almost surely, respectively, but the statements are weaker than those obtained in Theorem 1. The qualification "almost all x" refers to all Lebesgue points of f. It cannot be dropped because a density f is only defined up to a set of zero Lebesgue measure. Theorem 2 can basically not be improved because the conditions on h in A, B, and C are necessary for the standard kernel estimate: this result was first established by Deheuvels (1974) under various regularity conditions on K, f, and h. This is a fine occasion to present Deheuvels' result: in the next section, we will state an extended version of it stripped of most regularity conditions. The proof is given in Section 5.

In Theorem 2, h is also allowed to depend upon x, but in that case no L_1 consistency results can be derived from it via Theorems 2.7 and 2.8.

Several results in the spirit of Theorem 2 are known for the uniform convergence of the automatic kernel estimate, for example, those of Wagner (1975), Devroye and Wagner (1980) and Deheuvels and Hominal (1980). In the last reference, a sketch is given of the proof of the following result: if $h \to 0$ almost surely, and $nh^d/(\log n) \to \infty$ almost surely, and if f is uniformly continuous and K is a Riemann integrable density, then $\sup_x |f_n(x) - f(x)| \to 0$ almost surely.

3. POINTWISE CONVERGENCE OF THE STANDARD KERNEL ESTIMATE

Definition. A sequence of positive numbers a_n is called *semimonotone* if there exists a constant $c > 0$ such that $a_{n+m} \geq ca_n$, all $m, n \geq 1$. Note that this implies that either $\liminf_{n\to\infty} a_n = \infty$ or $\sup_n a_n < \infty$.

THEOREM 3. *Let f_n be the standard kernel estimate with smoothing factor h depending upon n only and bounded density K with compact support.*

1. Weak Version. The following are equivalent:
   A. $f_n \to f$ in probability, almost all x, some f.
   B. $f_n \to f$ in probability, almost all x, all f.
   C. $h \to 0$ and $nh^d \to \infty$.
   D. $\int |f_n - f| \to 0$ in probability, some f.
   E. $\int |f_n - f| \to 0$ completely, all f.
2. Strong Version. Let K also be Riemann integrable, and let $nh^d/(\log\log n)$ be semimonotone. Then the following are equivalent:
   A. $f_n \to f$ almost surely, almost all x, some f.
   B. $f_n \to f$ almost surely, almost all x, all f.
   C. $h \to 0$ and $nh^d/(\log\log n) \to \infty$.
   The Riemann integrability of K is not needed for $A \Rightarrow C$, and the semimonotonicity condition is not used in the proof of $C \Rightarrow B$.
3. Complete Version. Let $nh^d/(\log n)$ be semimonotone. Then the following are equivalent:
   A. $f_n \to f$ completely, almost all x, some f.
   B. $f_n \to f$ completely, almost all x, all f.
   C. $h \to 0$ and $nh^d/(\log n) \to \infty$.
   The semimonotonicity condition is not needed for the implication $C \Rightarrow B$.

The inclusion of Theorem 3 in this book is partly motivated by the following observation: the standard kernel estimate can be strongly L_1 consistent (i.e., $\int |f_n - f| \to 0$ almost surely) while at the same time it is not pointwise convergent (almost sure sense) at almost all x. This happens, for example, for all f when nh^d is chosen as $\log\log\log n$, as $\sqrt{\log\log n}$, or as $c\log\log n$ for a positive constant c.

4. EXAMPLES OF AUTOMATIC KERNEL ESTIMATES

Most automatic kernel estimates fall into one of two categories: the first category houses all the estimates in which one considers the main term in the asymptotic expansion of some error criterion, and estimates all the unknown factors from the data. The second category groups the estimates in which h is obtained by minimizing some criterion (such as the maximum likelihood suitably modified) *directly*. The latter approach often requires more computations, and the theoretical analysis of the properties of the automatic estimate is usually more difficult too.

Estimates Based on Asymptotic Expansions. First and foremost among the estimates based on asymptotic expansions is the parametric method illustrated in Section 5.6 for the L_1 error. The parametric method should give excellent results if the hypothesis that f belongs to the given parametric family of densities is correct. For example, if the parametric family is normal with unknown mean μ and variance σ^2, and if σ^2 is estimated from the data by $\hat{\sigma}^2$, then in the parametric method we use

$$h = \hat{\sigma}\left(\frac{15e\sqrt{2\pi}}{8n}\right)^{1/5} = \frac{1.6644\cdots\hat{\sigma}}{n^{1/5}}.$$

For all f for which $\hat{\sigma} \to c$ almost surely for some constant c (this includes nearly all densities when $\hat{\sigma}$ is a reasonable data-based estimate), we have, by Theorem 1, $\int|f_n - f| \to 0$ almost surely. The same remains valid when $\hat{\sigma}$ remains almost surely bounded away from 0 and infinity (this is satisfied for all f when $\hat{\sigma}$ is a robust quantile-based estimate; see Section 5.6).

In fact, more is true. It is a good exercise to obtain conditions under which $\hat{\sigma} \to c$ almost surely (or weakly) implies that $E(\int|f_n - f|) \sim E(\int|f_n^* - f|)$, where f_n^* is the kernel estimate in which, for the formula of h_n, $\hat{\sigma}$ is replaced by c. Thus, not only do we have consistency, we also have information about the rate of convergence. We should note however that this rate is not optimal unless f indeed belongs to the parametric family. See also Section 7.

The parametric method was first developed in depth by Deheuvels (1977) for R^1 and the criterion $E(\int(f_n - f)^2)$. His development was based upon Rosenblatt's fundamental result that when K is a bounded symmetric density and f is a bounded density with two continuous derivatives all in L_2, then the standard kernel estimate satisfies

$$E\left(\int(f_n - f)^2\right) \sim (nh)^{-1}\int K^2 + \frac{1}{4}h^4\left(\int x^2K(x)\,dx\right)^2\int f''^2 \qquad (2)$$

provided that $h \to 0$ and $nh \to \infty$ (Rosenblatt, 1956, 1971). From (2) it appears that the best value for h is given by $h = [A/n\int f''^2]^{1/5}$, where $A = \int K^2/(\int x^2K(x)\,dx)^2$ is a factor depending upon K only. The only unknown in this expression is $\int f''^2$, which, for the normal density, equals $3/(8\sqrt{\pi}\,\sigma^5)$. See also Deheuvels and Hominal (1980) for further discussions.

For the histogram estimate, the parametric method was developed by Scott (1979).

Others have proposed to estimate the unknown factor $\int f''^2$ in (2) from the data by nonparametric means. This leads to a two-step procedure: first, $\int f''^2$ is estimated, and then f is estimated by using the estimate of $\int f''^2$ in

$h = [A/nJf''^2]^{1/5}$ (Woodroofe, 1970; Nadaraya, 1974; Scott et al., 1977; Deheuvels and Hominal, 1980; Scott and Factor, 1981). It seems straightforward to mimick this work for the L_1 error. In both cases, however, we are again faced with the choice of parameters for nonparametric estimates (thus creating a new equally difficult problem as the one we tried to solve), and we assume that quantities such as Jf''^2 can be estimated consistently if at all. The parametric method is more robust in this respect. Finally, all the given methods are based upon some strong assumptions regarding f that may or may not be satisfied. For example, if the L_1 error is used, we need to estimate $J\sqrt{f}$ and $J|f''|$ (or a suitable generalization of it): the first term is infinite for the Cauchy density, while the second term is infinite for the uniform density. Large classes of important densities have to be excluded from further consideration, and this should be avoidable.

Heuristic Estimates. Among the estimates that are not based upon asymptotic expansions, the approaches of Wagner (1975), Silverman (1978), and others fall into a separate category. No attempt is made to attain some theoretically predicted performance (as for estimates that are based upon asymptotic expansions), or to optimize some criterion as we will illustrate below.

For example, Wagner (1975) considers $D_{n1}, \ldots, D_{nn}$, the distances between $X_1, \ldots, X_n$ and their respective kth nearest neighbors, where $k = \lfloor n^a \rfloor$, $0 < a < 1$. He suggests several schemes for choosing h such as

(i) h is chosen at random from $D_{n1}, \ldots, D_{nn}$.

(ii) $h = \sum_{i=1}^{n} (D_{ni}/n)$.

(iii) $h = \max_i D_{ni}$.

(iv) $h = \min_i D_{ni}$.

The number of possibilities is virtually unlimited. For (i) he has shown that for all f, $h \to 0$ almost surely, and $n^b h^{2d} \to \infty$ almost surely when $b > 1 - a$. Thus, for the kernels considered in Theorem 2 and for all $f, f_n \to f$ almost surely for almost all x, and $J|f_n - f| \to 0$ almost surely.

Optimizing Some Criterion. If h is chosen so that some criterion is maximized, we can hope to obtain an estimate that can be trusted for all (unknown) f, even for quite small sample sizes. The first and foremost criterion is based upon the *maximum likelihood* (ML) principle. It was first suggested for use in kernel density estimates by Duin (1976, paper submitted in 1973) and Habbema et al. (1974). See also a recent survey by Rudemo (1982). They suggest choosing h so as to maximize the likelihood

$$L(h) = \prod_{i=1}^{n} f_{ni}(X_i), \tag{3}$$

where

$$f_{ni}(x) = \frac{1}{(n-1)h^d} \sum_{\substack{j=1 \\ j \neq i}}^{n} K\left(\frac{x - X_j}{h}\right), \tag{4}$$

and to use this h in the standard kernel estimate. This cross-validated kernel density estimate (and a similarly defined cross-validated histogram estimate) seems to work well in most, but not all, situations. The cross-validation was needed in the first place because $\prod_{i=1}^{n} f_n(X_i)$ is maximal for $h = 0$.

The major difficulty with cross-validated estimates is to establish their consistency. Our Theorem 1 allows us to do this if we can prove something about the smoothing factor h. The first occurrence of a consistency proof for the cross-validated kernel estimate in R^1 is in Chow et. al. (1983). Because of its practical importance and because of the novel technique of proof, we will state their main result here, suitably generalized to R^d and stripped of unnecessary conditions on f. The proof is given in Section 5.

THEOREM 4. *Choose h such that $L(h) \geq a \sup_{h>0} L(h)$ for some $a \in (0,1)$. Let f_n be the cross-validated kernel density estimate, and let f be a density with compact support. Let K be a bounded compact support kernel with $K \geq cI_{S_{0,r}}$ for some $c, r > 0$ (recall that $S_{x,r}$ is the closed sphere of radius r centered at x), and assume that K is Riemann integrable. Then $h \to 0$ and $\liminf_{n\to\infty} nh^d/(\log n) > 0$ almost surely.*

Thus, (as a corollary of Theorem 2) $f_n \to f$ almost surely for almost all x, and $\int |f_n - f| \to 0$ almost surely.

Sometimes solutions of some optimization methods are truncated between two bounds to "force" consistency. The strength of Theorem 4 is that no such bounding is necessary. The only requirement on f is that it have compact support. Consider however the following variation on R^1 of the method mentioned above:

Step 1. Apply a monotone transformation (such as $T(x) = x/(1 + |x|)$): $R \to [-1, 1]$ to the data, thus obtaining $Y_1, \ldots, Y_n$.

Step 2. Construct the cross-validated kernel density estimate on $[-1, 1]$. (This is *always* consistent, by Theorem 4.)

Step 3. Estimate f by the transformed cross-validated kernel density estimate.

Because the L_1 distance between densities is invariant under monotone transformations of the coordinate axes, the L_1 error equals the error committed in estimating the density of Y_1 by the cross-validated density

estimate in Step 2. In other words, the variation given here is consistent for *all* densities f.

Although Theorem 4 is reassuring, it does not tell us anything about the rate of decrease to 0 for $E(\int |f_n - f|)$, and this seems to be a challenging open problem. Also, we have not explained why some strong tail condition on f appears in the statement of Theorem 4. Take a density f on R^1, let K be 0 outside $[-1, 1]$, and let $X_{(1)} < \cdots < X_{(n)}$ be the order statistics of $X_1, \ldots, X_n$. If $h < X_{(n)} - X_{(n-1)}$, we have $f_{ni}(X_i) = 0$, where $X_i = X_{(n)}$, and thus $L(h) = 0$. Thus, the h that is actually chosen by the cross-validation method satisfies for all n: $h \geq X_{(n)} - X_{(n-1)}$. Now, let K be a kernel bounded by M. For each $\varepsilon > 0$ we can find $\delta > 0$ such that

$$\int_{\delta f > 2M} f \geq \varepsilon > 0.$$

Now,

$$\int |f_n - f| = 2\int_{f > f_n} f - f_n \geq 2\int_{f > 2f_n} \frac{f}{2} \geq \int_{f > 2M/h} f$$

is at least equal to ε when $h \geq \delta$. We tacitly used the fact that $f_n \leq M/h$. Thus,

$$P\left(\int |f_n - f| \geq \varepsilon\right) \geq P(h \geq \delta) \geq P\left(X_{(n)} - X_{(n-1)} \geq \delta\right). \tag{5}$$

Therefore, convergence in probability to 0 is impossible for $\int |f_n - f|$ if $X_{(n)} - X_{(n-1)} \to \infty$ in probability. This fact was first pointed out by Schuster and Gregory (1981). For example, any density for which $\lim_{x \to \infty} f(x)/\int_x^\infty f(y)\,dy = 0$, $\int_x^\infty f > 0$ for all x (such as *all* the densities with a tail decreasing at a polynomial rate c/x^a, $a > 1$), must have $X_{(n)} - X_{(n-1)} \to \infty$ in probability. This class contains the Cauchy density, the Student's t densities, the Pareto densities, and all stable densities except the normal density.

In Schuster and Gregory (1981), some method is given to take care of the nonconsistency for long-tailed distributions. The borderline between consistency and nonconsistency seems to be the exponential distribution (for which $h \not\to 0$ in probability, and thus we have nonconsistency). Distributions with smaller tails seem to be safe. The condition that f have compact support (Theorem 4) is too strong in this respect.

Experimental evidence is given in Scott and Factor (1981), Rudemo (1982), Schuster and Gregory (1981). Chow et al. (1983) give a theorem in

the spirit of Theorem 4 for cross-validated histogram estimates. Geman (1981) and Geman and Hwang (1982) apply the ML principle to a kernel estimate in which not the X_i's but some design variables $x_1, \ldots, x_n$ are chosen as the centers for the kernels. Schuster and Gregory (1978) cut the sample artificially in two, determine h by maximizing

$$\prod_{i=1}^{n/2} f_{n/2}(X_i), \tag{6}$$

where $f_{n/2}(x) = (2/n)\sum_{j=n/2+1}^{n} h^{-d}K((x - X_j)/h)$, and they use this h in the original kernel estimate. This approach requires less computational effort, but seems to give poorer results than if cross-validation were used.

Finally, Hall (1982a, b) gives evidence that the cross-validation method when used for f that are concave on $[0, 1]$ yields smoothing factors h that are of the order of magnitude $n^{-1/3}$ (which is necessarily suboptimal in certain cases).

The cross-validation method can also be used on other criteria besides the maximum likelihood criterion. For example, in an attempt to find the h that minimizes $\int (f_{nh} - f)^2$, where f_{nh} is the kernel estimate with smoothing factor h and kernel K, Hall (1983a, b), Rudemo (1982), and Bowman (1982) suggested a minimization of $\int f_{nh}^2 - 2M_{nh}$, where M_{nh} is a sample-based cross-validation estimate of $\int f_{nh} f$, for example,

$$M_{nh} = \frac{1}{n}\sum_{i=1}^{n} f_{nhi}(X_i),$$

where

$$f_{nhi}(x) = \frac{1}{n-1}\sum_{j \neq i} h^{-1}K\left(\frac{x - X_j}{h}\right), \qquad i = 1, \ldots, n.$$

Stone (1984) observed that

$$M_{nh} = \frac{1}{n(n-1)}\sum_{i \neq j} K_h(X_i - X_j), \qquad \int f_{nh}^2 = \frac{1}{n^2}\sum_{i,j}(K^*K)_h(X_i - X_j).$$

The value h^* thus obtained is best possible in the following sense:

$$\frac{\int (f_{nh^*} - f)^2}{\min_h \int (f_{nh} - f)^2} \to 1 \quad \text{almost surely}$$

for all bounded f on R whenever K is Lipschitz (with any positive Lipschitz power), K is symmetric, K has compact support, and $\int K^2 < 2K(0)$ (Stone, 1984). K does not have to be nonnegative, as long as $\int K = 1$. For an analogous result for the histogram estimate, see Stone (1983). It is not known whether the given cross-validation estimate is consistent for all f. As we have seen, minimizing the L_2 error could actually lead to extremely poor L_1 rates of convergence. The question still remains of the construction of an estimator h^* such that

$$\frac{\int |f_{nh^*} - f|}{\min_h \int |f_{nh} - f|} \to 1 \quad \text{almost surely}$$

for all f, or with "$\to 1$" replaced by "$\leq C + o(1)$."

5. PROOFS

For the proof of Theorem 1, a few key Lemmas are needed:

LEMMA 1. *For any density K on R^d,*

$$\lim_{h \to 1} \int |K_h - K| = 0,$$

where $K_h(x) = h^{-d}K(x/h)$.

Proof. When K is continuous, the statement is obviously true: $K_h \to K$ at all x. Thus, by Scheffé's Theorem 2.7, we are done.

When K is an arbitrary density, we can find for every $\varepsilon > 0$ a continuous density $\tilde{K}$ such that $\int |K - \tilde{K}| < \varepsilon$. But because

$$\int |K_h - K| \leq \int |K_h - \tilde{K}_h| + \int |\tilde{K}_h - \tilde{K}| + \int |\tilde{K} - K|,$$

we have

$$\limsup_{h \to 1} \int |K_h - K| \leq 2\varepsilon + \lim_{h \to 1} \int |\tilde{K}_h - \tilde{K}| = 2\varepsilon,$$

and Lemma 1 is proved.

For an arbitrary density K introduce the notation

$$\phi(\delta) = \sup_{1-\delta<h<1+\delta} \int |K_h - K|.$$

For any $\delta \in (0,1)$ we have $0 \le \phi(\delta) \le 2$. Also, by Lemma 1, $\lim_{\delta \downarrow 0} \phi(\delta) = 0$. At this point we would like to make the dependence of f_n upon h explicit, and we will write f_{nh} instead of f_n.

LEMMA 2. *Consider a sequence of intervals $H_n = [h_n', h_n'']$, where $h_n'' \to 0$ and $nh_n'^d \to \infty$. Then, for each $\varepsilon > 0$ there exist positive numbers n_0 and r such that*

$$P\left(\sup_{h \in H_n} \int |f_{nh} - f| > \varepsilon \right) \le \exp(-rn\varepsilon^2), \quad \text{all } n \ge n_0.$$

Proof. By our assumptions,

$$\lim_{n \to \infty} \frac{h_n''^d}{nh_n'^d} = 0,$$

and thus

$$1 \le \frac{h_n''}{h_n'} = a_n n^{1/d},$$

where $\lim_{n \to \infty} a_n = 0$. Let $\delta_n \ge 0$ be the solution of the equation

$$(1 + \delta_n)^n = a_n n^{1/d}.$$

Clearly, $\delta_n \to 0$. Next, introduce

$$h_{ni} = (1 + \delta_n)^i h_n', \qquad i = 0, 1, 2, \ldots, n.$$

Thus, $h_{n0} = h_n'$ and $h_{nn} = h_n''$, so that

$$\sup_{h \in H_n} \int |f_{nh} - f| \le \sup_{1 \le i \le n} \left(\int |f_{nh_{ni-1}} - f| + \sup_{h_{ni-1} \le h < h_{ni}} \int |f_{nh} - f_{nh_{ni-1}}| \right). \tag{7}$$

For each $u > h$,

$$\int |f_{nh} - f_{nu}| \le \frac{1}{n} \sum_{i=1}^{n} \int |K_h(x - X_i) - K_u(x - X_i)|\, dx$$

$$= \int \left| \left(\frac{u}{h}\right)^d K\left(\frac{u}{h}x\right) - K(x) \right| dx \le \phi\left(\frac{u}{h} - 1\right).$$

Thus,

$$\sup_{h_{ni-1} \le h \le h_{ni}} \int |f_{nh} - f_{nh_{ni-1}}| \le \sup_{h_{ni-1} \le h \le h_{ni}} \phi\left(\frac{h}{h_{ni-1}} - 1\right)$$

$$\le \phi\left(\frac{h_{ni}}{h_{ni-1}} - 1\right) = \phi(\delta_n). \tag{8}$$

There exists for each $\varepsilon > 0$ a number $n_1 > 0$ such that $\phi(\delta_n) < \varepsilon, n \ge n_1$. Thus, in view of (7) and (8),

$$P\left(\sup_{h \in H_n} \int |f_{nh} - f| > 2\varepsilon \right) \le \sum_{i=1}^{n} P\left(\int |f_{nh_{ni-1}} - f| > \varepsilon \right). \tag{9}$$

We will now apply Remark 3.1: each term on the right-hand side of (9) is bounded from above by $\exp(-rn\varepsilon^2), n > n_0$, where r and n_0 are positive constants, provided that for each $i \in \{1, \ldots, n\}$, we have

$$\left(\frac{c_0(\varepsilon)}{n}\right)^{1/d} \le h_{ni} \le h_0(\varepsilon)$$

in the notation of Remark 3.1. But this is satisfied when $h_n'' < h_0(\varepsilon)$ and $(c_0(\varepsilon)/n)^{1/d} < h_n'$. This concludes the proof of Lemma 2 since for all n large enough, (9) is bounded from above by $n \exp(-rn\varepsilon^2)$.

Proof of Theorem 1. We prove the complete convergence. The strong and weak convergence can be obtained similarly by substituting the word "completely" by the words "almost surely" or "in probability". Assume that for each $\varepsilon > 0$, we have $I_{[h + 1/(nh^d) \ge \varepsilon]} \to 0$ completely. This is equivalent to the condition that there exists $\varepsilon_n' \downarrow 0$ such that $I_{[h+1/(nh^d) \ge \varepsilon_n']} \to 0$ completely. Now, define

$$\varepsilon_n = \max\left(\varepsilon_n', \left(\frac{1}{n}\right)^{1/(d+1)}\right),$$

and verify that $\varepsilon_n \to 0$ and $I_{[h+1/(nh^d)>\varepsilon_n]} \to 0$ completely. Next, define $H_n = [h_n', h_n'']$ by

$$h_n' = (n\varepsilon_n)^{-1/d},$$

$$h_n'' = \varepsilon_n.$$

By definition of ε_n, we have $h_n' \le h_n''$. Furthermore, $h_n'' \to 0$ and $nh_n'^d \to \infty$. Thus, because

$$I_{[h+1/(nh^d)\ge\varepsilon_n]} = \tfrac{1}{2}\Big(I_{[h+1/(nh^d)\ge h_n'']} + I_{[h+1/(nh^d)\ge (nh_n'^d)^{-1}]}\Big)$$

$$\ge \tfrac{1}{2}\Big(I_{[h\ge h_n'']} + I_{[h\le h_n']}\Big)$$

$$= \tfrac{1}{2} I_{[h\notin H_n]},$$

we see that $I_{[h\notin H_n]} \to 0$ completely. Theorem 1 now follows from this observation, Lemma 2, and the inequality, valid for all $\varepsilon > 0$:

$$I_{[|f_n - f|>\varepsilon]} \le I_{[h\notin H_n]} + I_{[\sup_{h\in H_n}|f_{nh}-f|>\varepsilon]}.$$

The proof of Theorem 1 illustrates very nicely the power and depth of Theorem 3.1: we required hardly any new technical tools. For the proofs of Theorems 2 and 3, quite a few new elements are needed. Until we complete the proof of Theorem 3 we will assume that K is a density bounded by K^* with support contained in $S_{0,c}$, the closed sphere of radius c centered at 0.

LEMMA 3 (Convergence of the Bias). *Let h_n'' be a sequence of positive numbers tending to 0 as $n \to \infty$. For all densities f on R^d we have*

$$\lim_{n\to\infty} \sup_{0<h\le h_n''} |f * K_h - f| = 0, \quad \textit{almost all } x.$$

Proof. We can bound the said supremum from above by

$$\sup_{0<h\le h_n''} \int_{S_{0,ch}} |f(x-y) - f(x)| K_h(y)\, dy$$

$$\le K^* \sup_{0<h\le h_n''} \left(\int_{S_{0,ch}} \frac{|f(x-y)-f(x)|dy}{\lambda(S_{0,ch})} \right) \lambda(S_{0c}),$$

where λ is Lebesgue measure. By the Lebesgue density Theorem 2.2, this tends to 0 for almost all x.

LEMMA 4. *For every nonnegative Riemann integrable function K bounded by K^* on $[0,1]^d$, and for every $\varepsilon > 0$, there exists an integer N and nonnegative numbers $a_i \in [0, K^*], 1 \le i \le N^d$, such that the function*

$$K_1(x) = \sum_{i=1}^{N^d} a_i I_{A_i}(x), \qquad x \in [0,1]^d,$$

in which the A_i's are the rectangles formed by the products of intervals of the form $[(j-1)/N, j/N)$, $1 \le j < N$, or $[(N-1)/N, 1]$, satisfies:

(i) $|K_1(x) - K(x)| < \varepsilon$, all $x \notin A$ = union of some A_i's;

(ii) $0 \le K_1(x) \le K^*$, all x;

(iii) $\lambda(A) < \varepsilon$.

LEMMA 5 (Fundamental Inequalities for the Uniform Deviation). *Let $\varepsilon > 0$ be an arbitrary number, let x be a Lebesgue point for f (see Theorem 2.2), and let h'_n and h''_n be two positive number sequences satisfying $0 < h'_n \le h''_n \downarrow 0$. Let f_{nh} be the estimate (1) with smoothing factor h. Then*

$$\sup_{h'_n \le h \le h''_n} P\big(|f_{nh}(x) - f * K_h(x)| \ge \varepsilon\big) \le 2\exp\big(-bnh_n'^d\big),$$

where b can be taken as $\varepsilon^2/(2K^*(f(x) + o(1) + \varepsilon))$.

If K is Riemann integrable, then also

$$P\Big(\sup_{h'_n \le h \le h''_n} |f_{nh}(x) - f * K_h(x)| \ge \varepsilon\Big) \le \frac{a\exp\big(-bnh_n'^d\big)}{1 - \exp\big(-b'nh_n'^d\big)}$$

for some positive constants a, b, b' not depending upon n.

Proof. Bennett (1962) has shown that for independent identically distributed zero mean random variables Z_i with $|Z_i| \le t$, and for all $\varepsilon > 0$,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right| > \varepsilon\right) \le 2\exp\left(-\frac{n}{2t}\left(\left(1 + \frac{\sigma^2}{2t\varepsilon}\right)\log\left(1 + \frac{2t\varepsilon}{\sigma^2}\right) - 1\right)\right)$$

$$\le 2\exp\left(\frac{-n\varepsilon^2}{2(\sigma^2 + t\varepsilon)}\right),$$

where $\sigma^2 = E(Z_1^2)$. The last inequality follows from $\log(1+u) \ge 2u/(2+u)$, valid for all $u > 0$.

Our first inequality follows by replacing Z_i by $h^{-d}(K((x - X_i)/h) - E(K((x - X_i)/h)))$, which is bounded in absolute value by $t = K^*/h^d$, and has variance $\sigma^2 \le K^* f * K_h(x)/h^d = K^*(f(x) + o(1))/h^d$ uniformly on $[0, h''_n]$ (by Lemma 3).

For the second inequality, we take a positive number δ (to be specified later), and let $h_{ni} = h'_n(1+\delta)^i$, $i \geq 0$. Let i_0 be such that $h_{ni_0-1} \leq h''_n < h_{ni_0}$. We have, if we write estimate (1) as $f_{nh}(x)$ to make the dependence upon h explicit,

$$\begin{aligned}
&\sup_{h'_n \leq h \leq h''_n} |f_{nh}(x) - f * K_h(x)| \\
&\leq \sup_{0<i\leq i_0} \Bigg[|f_{nh_{n,i-1}}(x) - f * K_{h_{n,i-1}}(x)| \\
&\qquad + \sup_{h_{n,i-1} \leq h, h' \leq h_{ni}} |f * K_h(x) - f * K_{h'}(x)| \\
&\qquad + \sup_{h_{n,i-1} \leq h, h' \leq h_{ni}} |f_{nh}(x) - f_{nh'}(x)| \Bigg] \\
&= \sup_{0<i\leq i_0} [U_i + V_i + W_i].
\end{aligned}$$

By the first part of Lemma 5, for $\varepsilon > 0$,

$$P(U_i \geq \varepsilon) \leq 2\exp\left(\frac{-nh_{n,i-1}^d \varepsilon^2}{2K^*(f(x) + \varepsilon + o(1))} \right),$$

where the $o(1)$ terms does not depend upon i (since $0 \leq i \leq i_0$). By Lemma 3,

$$\sup_{0<i\leq i_0} V_i \leq 2 \sup_{h \leq h_{ni_0}} |f * K_h(x) - f(x)| \to 0.$$

For fixed $\varepsilon > 0$, find K_1, N, $a_1, \ldots, a_{N^d}$, and sets A_j as in Lemma 4 (after having replaced $[0,1]^d$ by $[-c,c]^d$). The set A also keeps its meaning from Lemma 4. We introduce the notation μ and μ_n for the measure induced by f, and the empirical measure defined by $X_1, \ldots, X_n$, respectively. Also, Δ is the difference operator between sets.

Without loss of generality, we can assume that all sets A_j are strictly contained in one quadrant, such as $[0,c]^d$. We need a few geometrical facts now. Let h, h' be numbers in the interval $[h_{n,i-1}, h_{ni}]$, and let A_j be fixed, for example, $A_j = [a_1, a'_1] \times \cdots \times [a_d, a'_d]$. Then, $(x + hA_j)\Delta(x + h'A_j) \subseteq (x + h_{ni}B_j)$, where B_j is a set of fixed form and dimensions determined by A_j, d, and δ only. Also, $\lambda(B_j) \leq 2c^d\delta$.

To prove this first geometrical fact, we need only show that $uA_j\Delta u'A_j \subseteq B_j$ for all $u, u' \in [1/(1+\delta), 1]$. First, take

$$B_j = [a_1, a_1'] \times \cdots \times [a_d, a_d'] - \left[a_1, \frac{a_1'}{1+\delta}\right] \times \cdots \times \left[a_d, \frac{a_d'}{1+\delta}\right]$$

$$+ \left[\frac{a_1}{1+\delta}, a_1'\right] \times \cdots \times \left[\frac{a_d}{1+\delta}, a_d'\right] - [a_1, a_1'] \times \cdots \times [a_d, a_d'],$$

where the $-$ operators are considered before the union operator $+$. We note that B_j is contained in $[-c, c]^d$. Also,

$$\lambda(B_j) \le \left(a_1'\left(1 - \frac{1}{1+\delta}\right) + a_1\left(1 - \frac{1}{1+\delta}\right)\right) a_2' a_3' \cdots a_d'$$

$$\le 2a_1' a_2' \cdots a_d' \delta \le 2c^d \delta.$$

Second, let A be the set of Lemma 4, that is, it is the union of M disjoint rectangles A_j, and let B be the set of all points contained in uB, where $u \in [1/(1+\delta), 1]$. Then, $B \subseteq [-c, c]^d$, and by the previous derivation for a single rectangle, $\lambda(B) \le \lambda(A) + \sum_{j=1}^{M} \lambda(B_j) \le \lambda(A) + 2Mc^d\delta$. We can now obtain the following crucial upper bound for W_i:

$$W_i \le \sup_{h_{n,i-1} \le h, h' \le h_{ni}} \sum_{j=1}^{N^d} \int a_j |I_{x-hA_j}(y) - I_{x+h'A_j}(y)| \frac{\mu_n(dy)}{h_{ni-1}^d}$$

$$+ 2 \sup_{h_{n,i-1} \le h \le h_{ni}} \sum_{j=1}^{N^d} \int_{x+hA_j} \left| K\left(\frac{x-y}{h}\right) - K_1\left(\frac{x-y}{h}\right)\right| \frac{\mu_n(dy)}{h_{n,i-1}^d}$$

$$\le \sum_{j=1}^{N^d} \frac{a_j \mu_n(x + h_{ni}A_j\Delta x + h_{n,i-1}A_j)}{h_{ni-1}^d}$$

$$+ \frac{2\varepsilon\mu_n(x + [-c, c]^d h_{ni})}{h_{ni-1}^d} + \frac{2K^*\mu_n(x + h_{ni}B)}{h_{n,i-1}^d}$$

$$\le h_{n,i-1}^{-d}\left(\sum_{j=1}^{N^d} K^*\mu_n(x + h_{ni}B_j) + 2\varepsilon\mu_n(x + [-c, c]^d h_{ni})\right.$$

$$\left. + 2K^*\mu_n(x + h_{ni}B)\right)$$

$$= \sum_{j=1}^{N^d} W_{ij} + W_i' + W_i'',$$

where μ_n is the empirical measure for $X_1, \ldots, X_n$. For a given $\eta > 0$, we find $\varepsilon, \delta > 0$ such that the expected value of each W_{ij} does not exceed $\eta/3N^d$, and the expected values of W_i' and W_i'' do not exceed $\eta/3$. This corresponds to the requirement that

$$(f(x) + o(1))(1 + \delta)^d K^* 2c^d \delta < \eta/(3N^d);$$

$$(f(x) + o(1))(1 + \delta)^d 2\varepsilon(2c)^d < \eta/3;$$

$$(f(x) + o(1))(1 + \delta)^d 2K^*(\varepsilon + 2Mc^d\delta) < \eta/3.$$

Once again, the $o(1)$ terms do not depend upon i, so that all three inequalities can be satisfied for all n large enough, uniformly in i. A small technical note is in order here: it seems necessary to choose ε first under the assumption that δ does not exceed 2. This fixes N and M, so that in a second step we can choose δ.

For each i, we have by simple bounding techniques,

$$P(W_i > 2\eta) \le \sum_{j=1}^{N^d} P\left(W_{ij} - E(W_{ij}) > \frac{\eta}{3N^d}\right)$$

$$+ P\left(W_i' - E(W_i') > \frac{\eta}{3}\right) + P\left(W_i'' - E(W_i'') > \frac{\eta}{3}\right). \tag{10}$$

Uniformly in i and j, we know that for all $n \ge n_0$, all expected values are smaller than $\eta/3$. Also, each of the W_{ij}'s, W_i''s, and W_i'''s can be written as $(1/n)\sum_{m=1}^n Y_m$, where the Y_m's are independent bounded nonnegative random variables with absolute value not exceeding r/h_{ni}^d, where

$$r = \max(2K^*, 2\varepsilon)(1 + \delta)^d.$$

Thus, by another application of Bennett's inequality, we see that each probability on the right-hand side of (10) does not exceed

$$2\exp\left(-\frac{n(\eta/3N^d)^2}{2\left((\eta/3)(r/h_{ni}^d) + (r/h_{ni}^d)(\eta/3N^d)\right)}\right) = 2\exp\left(-bnh_{ni}^d\right)$$

by definition of b. A combination of all the bounds derived above shows us

that for all n greater than some n_1,

$$P\Big(\sup_{h_n' \le h \le h_n''} |f_{nh}(x) - f^* K_h(x)| > 4\eta\Big)$$

$$\le \sum_{i=1}^{i_0} 2\exp\big(-snh_n'^d(1+\delta)^{d(i-1)}\big)$$

$$+ (N^d + 2)2\exp\big(-bnh_n'^d(1+\delta)^{di}\big), \tag{11}$$

where $s = \eta^2/4K^*(f(x) + \eta)$. The right-hand side of (11) is again bounded from above, albeit very crudely, by

$$\sum_{i=0}^{\infty} b' \exp\big(-b''nh_n'^d(1+\delta)^i\big)$$

$$\le \sum_{i=0}^{\infty} b' \exp\big(-b''nh_n'^d(1+\delta i)\big)$$

$$= \frac{b' \exp\big(-b''nh_n'^d\big)}{1 - \exp\big(-b''\delta nh_n'^d\big)}$$

for some positive constants b', b''. This concludes the proof of Lemma 5.

LEMMA 6 (A Binomial Tail Inequality). *Let Z be a binomial (n, p) random variable, with $p = p(n) \in (0,1)$ varying in such a way that $p + np^2 = o(1)$ but $\lim_{n\to\infty} np = \infty$. Then, for constant $\delta > 0$,*

$$P(Z - np \ge \delta np) \ge \frac{1 + o(1)}{\big(2\pi(1+\delta)^3 np\big)^{1/2}} \exp(-npH(\delta)),$$

where $0 < H(\delta) = (1+\delta)\log(1+\delta) - \delta \to 0$ as $\delta \downarrow 0$.

Proof. Let k be the ceiling function of $np(1+\delta)$. Then,

$$P(Z - np \ge \delta np) \ge \binom{n}{k} p^k (1-p)^{n-k} \ge \frac{(n-k+1)^k}{k!} p^k (1-p)^n e^{pk}$$

$$\ge \frac{(np)^k}{k!}(1-p)^n e^{pk}(1 - k^2/n).$$

Since $k^2 = o(n)$, $pk = o(1)$, and $k! \sim (k/e)^k\sqrt{2\pi k}$, the lower bound is

$$(1 + o(1))\left(\frac{npe}{k}\right)^k \frac{e^{-np}}{\sqrt{2\pi k}} = (1 + o(1))\frac{e^{k\ np}(1+\delta)^{-k}}{\sqrt{2\pi k}}$$

$$\geq (1 + o(1))\frac{e^{\delta np - k\log(1+\delta)}}{\sqrt{2\pi k}}$$

$$\geq (1 + o(1))\frac{e^{-npH(\delta)}}{(1+\delta)\sqrt{2\pi k}},$$

from which the sought inequality follows.

LEMMA 7 (Exponential Lower Bounds for Large Deviations). *Let f be an arbitrary density on R^d, and let x be a Lebesgue point of f with $f(x) > 0$. Let $\varepsilon > 0$ be a constant, and let $h = h_n$ be a sequence of positive numbers satisfying $h + nh^{2d} = o(1)$, $\lim_{n\to\infty} nh^d = \infty$. Let $H(\cdot)$ be defined as in Lemma 6, and let $\delta = 2\varepsilon/f(x)$. Then, for the kernel estimate* (1),

$$P(f_n(x) - f * K_h(x) \geq \varepsilon)$$

$$\geq \frac{1 + o(1)}{\left(2\pi nh^d f(x)(2c)^d(1+\delta)^3\right)^{1/2}}$$

$$\times \exp\left(-nh^d H(\delta)(f(x) + o(1))(2c)^d\right).$$

Proof. Let Y be a random vector defined as X restricted to $x + [-c, c]^d h$. Define

$$g_n(x) = \frac{1}{n}\sum_{i=1}^{n} h^{-d} K\left(\frac{x - Y_i}{h}\right),$$

where $Y_1, Y_2, \ldots$ are independent and distributed as Y. It is clear that $f_n(x)$ is distributed as $(N/n)g_N(x)$, where N is independent of the Y_i's, and distributed as the number of X_i's in $x + [-c, c]^d h$. Also, $E(f_n(x)) = pE(g_n(x))$, where $p = P(X_1 \in A = x + [-c, c]^d h) = (2c)^d h^d (f(x) +$

$o(1)$). We have the following inclusion, valid for all n large enough:

$$P\big(f_n(x) \ge E\big(f_n(x)\big) + \varepsilon\big)$$

$$\ge P\big(N \ge np(1+\delta)\big)$$

$$\times \inf_{k \ge np(1+\delta)} P\left(g_k(x) \ge E\big(g_k(x)\big) - \frac{\varepsilon}{2p(1+\delta)}\right). \qquad (12)$$

Indeed, on a rich enough probability space, we can think of $f_n(x)$ as being equal to $(N/n)g_N(x)$, where $Y_1, \ldots, Y_N$ is the subset of $X_1, \ldots, X_n$ that falls in A. If $N \ge np(1+\delta)$ and $g_N(x) \ge E\big(g_N(x)\big) - \varepsilon/2p(1+\delta)$, then

$$f_n(x) = \frac{N}{n} g_N(x) \ge \frac{np(1+\delta)}{n}\left(E\big(g_N(x)\big) - \frac{\varepsilon}{2p(1+\delta)}\right)$$

$$= p(1+\delta)\left(\frac{E\big(f_n(x)\big)}{p} - \frac{\varepsilon}{2p(1+\delta)}\right)$$

$$= E\big(f_n(x)\big) + \delta E\big(f_n(x)\big) - \frac{\varepsilon}{2}$$

$$\ge E\big(f_n(x)\big) + \varepsilon, \quad n \text{ large enough.}$$

This explains (12). By Chebyshev's inequality and the fact that $\mathrm{Var}(g_k(x)) \le K^*(f(x) + o(1))/kh^d p$, we see that (12) is at least equal to

$$P(N - np \ge \delta np) \inf_{k \ge np(1+\delta)} \left(1 - \left(\frac{2p(1+\delta)}{\varepsilon}\right)^2 \mathrm{Var}\big(g_k(x)\big)\right)$$

$$\ge P(N - np \ge \delta np)\left(1 - \frac{K^*(f(x) + o(1))(2p)^2(1+\delta)^2}{np(1+\delta)\varepsilon^2 h^d p}\right)$$

$$= P(N - np \ge \delta np)(1 - o(1)), \qquad (13)$$

to which Lemma 6 can be applied since N is binomial (n, p) with $p + np^2 = o(1)$ and $\lim_{n\to\infty} np = \infty$. This concludes the proof of Lemma 7.

At this point we would like to make the dependence of f_n upon h explicit, and we shall use the notation f_{nh}. A quantity of crucial importance to us will be

$$D_n(x) = \sup_{H_n} |f_{nh}(x) - f(x)|,$$

where the supremum is over all values of h in the interval $H_n = [h'_n, h''_n]$, and $0 < h'_n \le h''_n < \infty$ only depend upon n.

LEMMA 8. *Let K be a bounded Riemann integrable density with compact support, and let $h''_n = o(1)$. Then, for every density f on R^1:*

A. *If $nh'^d_n \to \infty$, then $D_n(x) \to 0$ in probability, almost all x.*

B. *If h'_n varies regularly with coefficient $r \le 0$ (i.e., $h'_{[tn]}/h'_n \to t^r$, all $t > 0$), and $nh'^d_n/\log\log n \to \infty$, then $D_n(x) \to 0$ almost surely, almost all x.*

C. *If $nh'^d_n/(\log n) \to \infty$, then $D_n(x) \to 0$ completely, almost all x.*

Proof. Parts A and C follow directly from Lemmas 3 and 5 and the trivial inequality

$$D_n(x) \le \sup_{H_n} |f_{nh}(x) - f * K_h(x)| + \sup_{H_n} |f * K_h(x) - f(x)|.$$

To prove statement B, we fix a small $\delta > 0$, and define a subsequence $n_i = \underline{(1+\delta)^i}$, $i = 0, 1, 2, \ldots$. Let

$$E_i = \sup_{n_i \le n < n_{i+1}} \sup_{h \in H_i^*} |f_{nh}(x) - f * K_h(x)|, \tag{14}$$

where

$$H_i^* = \left[\inf_{n_i \le n < n_{i+1}} h'_n, \sup_{n_i \le n < n_{i+1}} h''_n \right] = [h_i^*, h_i^{**}].$$

By Lemma 3, it is clear that

$$\sup_{n_i \le n < n_{i+1}} D_n(x) \le E_i + o(1) \quad \text{as } i \to \infty, \text{ all Lebesgue points } x.$$

Thus, to show that $D_n(x) \to 0$ almost surely for almost all x, it suffices to show that for all Lebesgue points, all $\varepsilon > 0$ and some $\delta(\varepsilon) > 0$,

$$\sum_{i=0}^{\infty} P(E_i > \varepsilon) < \infty$$

(by the Borel–Cantelli lemma). A simple bounding argument yields, for all

$n_i \le n < n_{i+1}$, and fixed h,

$$|f_{nh} - f * K_h| \le |f_{nh} - f_{n_i h}| + |f_{n_i h} - f * K_h|$$

$$\le \left(\frac{1}{n_i} - \frac{1}{n_{i+1}}\right) \sum_{j=1}^{n_i} K_h(x - X_j)$$

$$+ \frac{1}{n_i} \sum_{j=n_i+1}^{n_{i+1}} K_h(x - X_j) + |f_{n_i h} - f * K_h|$$

$$\le (\delta + o(1))\left(f_{n_i h} + \tilde{f}_{n_{i+1}-n_i, h}\right) + |f_{n_i h} - f * K_h|$$

$$\le (1 + \delta + o(1))|f_{n_i h} - f * K_h| + (\delta + o(1))|\tilde{f}_{n_{i+1}-n_i h} - f * K_h|$$

$$+ (\delta + o(1))2f * K_h. \tag{15}$$

Here $\tilde{f}_{nh}$ is an estimate independent of f_{nh} but distributed as f_{nh}. It is clear that E_i is not greater than the right-hand side of (15), preceded by $\sup_{h \in H_i^*}$. Since $h_i^{**} \to 0$ as $i \to \infty$, the last term in the upper bound is $2\delta f(x) + o(1)$ (Lemma 3). Now, for fixed $\varepsilon > 0$, let us choose δ so small that $\delta \le \frac{1}{2}, 2\delta f(x) < \varepsilon/4$, and i are so large that all the $o(1)$ terms in (15) do not exceed $\frac{1}{2}$ and the $o(1)$ term in $2\delta f(x) + o(1)$ does not exceed $\varepsilon/12$ (thus, the entire term does not exceed $\varepsilon/3$). For such large i, we have

$$E_i \le 2 \sup_{h \in H_i^*} |f_{n_i h} - f * K_h| + \sup_{h \in H_i^*} |\tilde{f}_{n_{i+1}-n_i h} - f * K_h| + \frac{\varepsilon}{3}. \tag{16}$$

By Lemma 5, there exist positive constants a, a', a'', b, b', b'' such that the probabilities that the first and second terms on the right-hand side of (16) exceed $\varepsilon/3$ do not exceed

$$\frac{a \exp\left(-a' n_i h_i^{*d}\right)}{1 - \exp\left(-a'' n_i h_i^{*d}\right)}$$

and

$$\frac{b \exp\left(-b'(n_{i+1} - n_i) h_i^{*d}\right)}{1 - \exp\left(-b''(n_{i+1} - n_i) h_i^{*d}\right)}, \tag{17}$$

respectively. The constants do not depend upon i.

For every $M > 0$, we can find i large enough such that $j \geq i$ implies $n_j h_j^{*d} \geq M \log\log n_j \geq M \log(j \log(1+\delta))$. For $j \geq i$, the bounds in (17) are smaller than

$$\frac{a + o(1)}{(j \log(1+\delta))^{Ma'}} \quad \text{and} \quad \frac{b + o(1)}{(j \log(1+\delta))^{Mb'\delta}}, \tag{18}$$

respectively. But both expressions in (18) are summable in j when $Ma' > 1$ and $Mb'\delta > 1$. This shows that $\delta(\varepsilon) > 0$ can be found such that

$$\sum_{i=0}^{\infty} P(E_i > \varepsilon) < \infty, \quad \text{all } \varepsilon > 0, \text{ all Lebesgue points of } f.$$

This concludes the proof of Lemma 8.

Proof of Theorem 2. Theorem 2 is based upon the inequality

$$|f_n(x) - f(x)| \leq \sup_{H_n} |f_{nh}(x) - f(x)| + \infty \cdot I_{[h_n \notin H_n]}, \tag{19}$$

where I is the indicator function of an event, and $\infty \cdot 0$ is 0. The integral versions follow from the pointwise versions (statements A and B) after noting that f_n is a density on R^d for each n, and that weak and strong extensions of Scheffé's Theorem are applicable (Theorem 2.8).

The proofs of the pointwise parts proceed by construction of a proper sequence $H_n = [h_n', h_n'']$. They are based upon increasing subsequences of the integers n_k' and n_k'', respectively. In all cases (A, B, and C), we have $n_1' = n_1'' = 1$. Also, $h_n'' = 1/k$ on $[n_k'', n_{k-1}'') \to 0$ as $k \to \infty$. Finally, h_n' and h_n'' are arbitrarily defined on $[n_1', n_2')$ and $[n_1'', n_2'')$, respectively.

Part A. Let

$$n_k'' = \inf\left(n: n > n_{k-1}'', \sup_{m \geq n} P\left(h_m \geq \frac{1}{k}\right) \leq \frac{1}{k}\right), \qquad k \geq 2,$$

$$n_k' = \inf\left(n: n > n_{k-1}', \sup_{m \geq n} P\left(mh_m^d \leq k\right) \leq \frac{1}{k}\right), \qquad k \geq 2,$$

$$h_n' = (k/n)^{1/d} \text{ on } [n_k', n_{k+1}'), \qquad k \geq 2.$$

Clearly, $nh_n'^d \to \infty$. Also, on $[n_k'', n_{k+1}'')$, $P(h_n \geq h_n'') = P(h_n \geq 1/k) \leq 1/k \to 0$ as $k \to \infty$. Similarly, $P(nh_n^d \leq nh_n'^d) = P(nh_n^d \leq k) \leq 1/k$ on $[n_k', n_{k+1}')$, and this tends to 0 as $k \to \infty$. This completes the proof of part A (apply (19) and Lemma 8).

Part C. Let

$$n_k'' = \inf\left(n:\ n > n_{k-1}'',\ \sum_{m \geq n} m^k P\left(h_m \geq \frac{1}{k} \right) \leq 2^{-k} \right), \qquad k \geq 2,$$

$$n_k' = \inf\left(n:\ n > n_{k-1}',\ \sum_{m \geq n} m^k P\left(\frac{m h_m^d}{\log m} \leq k \right) \leq 2^{-k} \right), \qquad k \geq 2,$$

$$h_n' = \left(k \frac{\log n}{n} \right)^{1/d} \text{ on } [n_k', n_{k+1}'), \qquad k \geq 2.$$

Clearly, $nh_n'^d/(\log n) \to \infty$. Also,

$$\sum_{n=1}^{\infty} P(h_n \geq h_n'') \leq n_2'' + \sum_{k \geq 2} \sum_{n=n_k''}^{n_{k+1}''-1} n^k P\left(h_n \geq \frac{1}{k} \right)$$

$$\leq n_2'' + \sum_{k \geq 2} 2^{-k} < \infty.$$

By an identical argument,

$$\sum_{n=1}^{\infty} P(h_n \leq h_n') \leq n_2' + \sum_{k \geq 2} 2^{-k} < \infty.$$

Thus, $\sum P(h_n \notin H_n) < \infty$ and, therefore, the right-hand side of (19) tends to 0 completely in view of Lemma 8.

Part B. Let

$$n_k'' = \inf\left(n:\ n > n_{k-1}'', P\left(\bigcup_{m \geq n} \left[\frac{m h_m^d}{\log \log m} \leq k \right] \right) \leq 2^{-k} \right), \qquad k \geq 2,$$

$$n_k' = \inf\left(n:\ n > n_{k-1}', P\left(\bigcup_{m \geq n} \left[h_m \geq \frac{1}{k} \right] \right) \leq 2^{-k} \right), \qquad k \geq 2,$$

$$h_n' = (k(\log \log n)/n)^{1/d} \text{ on } [n_k', n_{k+1}'), \qquad k \geq 2.$$

Check that $nh_n'^d/(\log \log n) \to \infty$, and that $h_n \geq h_n''$ finitely often almost

surely because on $[n_k'', n_{k+1}'')$,

$$P\left(\bigcup_{m\geq n}[h_m \geq h_m'']\right) \leq \sum_{j=k}^{\infty} P\left(\bigcup_{m=n_j''}^{n_{j+1}''}\left[h_m \geq \frac{1}{j}\right]\right)$$

$$\leq \sum_{j=k}^{\infty} 2^{-j} = 2^{-k+1} \to 0 \text{ as } k \to \infty.$$

In a similar way, it can be checked that $h_n \leq h_n'$ finitely often almost surely. Part B will be complete if we can find a sequence of positive numbers $h_n^* \leq h_n'$ such that $nh_n^{*d}/(\log\log n) \to \infty$ and that h_n^* is *regularly varying*. Lemma 8 and (19) will then complete the proof. The sequence $\phi(n) = nh_n'^d/(\log\log n)$ is nondecreasing by construction, and it tends to ∞. Define $\phi(t)$ on the real line by linear interpolation from $\phi(n)$. We will attempt to find a function $\psi(t)$ with $0 \leq \psi \leq \phi, \psi(t) \uparrow \infty$ as $t \uparrow \infty$, and $t\psi'(t)/\psi(t) \to 0$ as $t \to \infty$. This function ψ is thus slowly varying (Seneta, 1976, pp. 6–7). Then, we define $h_n^* = (\psi(n)(\log\log n)/n)^{1/d}$, and note that it satisfies all our requirements.

The function ψ that we suggest is continuous and piecewise linear with knots at $t_1 < t_2 < \cdots$, where $t_k \to \infty$. Let $t_1 = 1$, and set $\psi(t) = \phi(t)$ on $[0, 1]$. Given t_k and $\psi(t_k)$ we define t_{k+1} and $\psi(t_{k+1})$ as follows:

$$\psi(t_{k+1}) = \min\left(\phi(t_k), \psi(t_k)\left(1 + \frac{1}{2\log k}\right)\right),$$

$$t_{k+1} = \inf\left(t: t \geq t_k + 1, t/t_k \geq \frac{\psi(t_{k+1})}{\psi(t_k)},\right.$$

$$\left. t - t_k \geq \frac{(\psi(t_{k+1}) - \psi(t_k))t\log k}{\psi(t_{k+1})}\right).$$

Note that $t_k \geq k \to \infty$ as $k \to \infty$, that $\psi(t)/t \downarrow$, and that on $[t_k, t_{k+1})$,

$$\psi'(t) \leq \frac{\psi(t_{k+1})}{t_{k+1}\log k} \leq \frac{\psi(t)/t}{\log k},$$

The existence of t_{k+1} follows from the fact that we can always find $t \geq t_k + 1$ such that

$$t \geq \frac{t_k}{1 - \log k(1 - \psi(t_k)/\psi(t_{k+1}))},$$

because the denominator in the last expression is always at least $\frac{1}{2}$ (in other words, $t \geq 2t_k$ will always satisfy the given condition). Finally, $0 \leq \psi \leq \phi$ and $\psi(t) \uparrow \infty$ because $\phi(t) \to \infty$ and

$$\prod_{k=2}^{\infty} \left(1 + \frac{1}{2 \log k}\right) = \infty.$$

Proof of Theorem 3. Part 1 requires no new proof. The equivalence of C, D, and E is established in Theorem 3.1. Obviously, C $\Rightarrow$ B $\Rightarrow$ A by a combination of Lemmas 3 and 5.

Finally, A $\Rightarrow$ D by Glick's extension of Scheffé's Theorem (Theorem 2.8).

Part 3 is partially shown in Lemmas 3 and 5 (i.e., C $\Rightarrow$ B $\Rightarrow$ A). To show A $\Rightarrow$ C, we note that the necessity of $h_n = o(1)$ follows from part 1, and that the necessity of $nh^d/\log n \to \infty$ follows from Lemma 7: indeed,

$$\sum_{n=1}^{\infty} P\big(f_n(x) - E\big(f_n(x)\big) > \varepsilon\big) < \infty, \quad \text{all } \varepsilon > 0, \text{ almost all } x,$$

$h_n = o(1)$, and $nh_n^d \to \infty$ (both consequences of part 1 of this theorem) imply that

$$\sum_{n=1}^{\infty} \min\Big(1, \big(nh_n^d\big)^{-1/2} \exp\big(-anh_n^d\big)\Big) < \infty, \quad \text{all } a > 0 \tag{20}$$

by Lemma 7, since we can restrict ourselves to Lebesgue points for f, with $f(x) > 0$. If $nh_n^d/\log n$ is bounded by M, then the sum in (20) is at least equal to

$$\sum_{n \geq e^{1/M}}^{\infty} (M \log n)^{-1/2} n^{-aM},$$

which is not summable for $a \leq 1/M$. But if $nh_n^d/\log n$ cannot remain bounded, then $\lim_{n \to \infty} (nh_n^d/\log n) = \infty$ by its semimonotonicity. Hence A $\Rightarrow$ C.

Part 2 is the only nontrivial part of the theorem. Clearly, B $\Rightarrow$ A. Also, C $\Rightarrow$ B by Theorem 2 when K is Riemann integrable. We will now show that Lemma 7 suffices to prove that A $\Rightarrow$ C. Fix a constant $a > 0$, and define the subsequence n_i by $\exp(ai \log i)$, $i \geq 1$. Note that $(n_{i+1} - n_i)/n_i \sim (ei)^a$. Assume that we can show that whenever $nh_n^d/\log\log n \leq M < \infty$, $h_n \to 0$, $nh_n^d \to \infty$, and x is a Lebesgue point of f with $f(x) > 0$, then

$$P\big(|f_{n_i}(x) - E\big(f_{n_i}(x)\big)| > \varepsilon \text{ infinitely often}\big) = 1 \tag{21}$$

for ε small enough. By the semimonotonicity of $nh_n^d/\log\log n$, we must have that $\lim_{n\to\infty}(nh_n^d/\log\log n) = \infty$, to avoid a contradiction. The necessity of $h_n = o(1)$ follows from part 1 of this theorem, as does the necessity of $\lim_{n\to\infty} nh_n^d = \infty$. We will thus show (21) under the stated conditions. We have

$$\begin{aligned}
&\left[|f_{n_i}(x) - E\big(f_{n_i}(x)\big)| > \varepsilon \text{ i.o.}\right] \\
&\quad \supseteq \left[|\tilde{f}_i(x) - E\big(\tilde{f}_i(x)\big)| > 2\varepsilon \text{ i.o.}\right] \\
&\quad \cap \left[\frac{n_i}{n_{i+1}}|f_i^*(x) - E\big(f_i^*(x)\big)| > \varepsilon \text{ f.o.}\right], \qquad (22)
\end{aligned}$$

where

$$\tilde{f}_i(x) = (n_{i+1} - n_i)^{-1} \sum_{j=n_i+1}^{n_{i+1}} \frac{K\big((x - X_j)/h_{n_{i+1}}\big)}{h_{n_{i+1}}^d},$$

$$f_i^*(x) = n_i^{-1} \sum_{j=1}^{n_i} \frac{K\big((x - X_j)/h_{n_{i+1}}\big)}{h_{n_{i+1}}^d}.$$

Implication (22) follows from the inequality

$$\begin{aligned}
|f_{n_{i+1}}(x) - E\big(f_{n_{i+1}}(x)\big)| \geq{}& \frac{n_{i+1} - n_i}{n_{i+1}}|\tilde{f}_i(x) - E\big(\tilde{f}_i(x)\big)| \\
&- \frac{n_i}{n_{i+1}}|f_i^*(x) - E\big(f_i^*(x)\big)|.
\end{aligned}$$

By Lemma 5,

$$\begin{aligned}
&P\left(\frac{n_i}{n_{i+1}}|f_i^*(x) - E\big(f_i^*(x)\big)| > \varepsilon\right) \\
&\quad \leq 2\exp\left(-n_i h_{n_{i+1}}^d \frac{\varepsilon^2 (n_{i+1}/n_i)^2}{2K^*\big(f(x) + \varepsilon + o(1)\big)}\right) \\
&\quad = 2\exp\left(-n_{i+1} h_{n_{i+1}}^d \frac{\big(\varepsilon^2 + o(1)\big)(ei)^a}{2K^*\big(f(x) + \varepsilon + o(1)\big)}\right)
\end{aligned}$$

which is summable in i for all $a, \varepsilon > 0$ (since $nh_n^d \to \infty$), so that by the Borel–Cantelli lemma, the last event in (22) has probability 1. By the independence of its component events, the middle event in (22) occurs with probability 1 if and only if

$$\sum_{i=1}^{\infty} P\left(|\tilde{f}_i(x) - E\left(\tilde{f}_i(x)\right)| > 2\varepsilon\right) = \infty. \tag{23}$$

A lower bound for the ith probability in (23) is given in Lemma 7 if we replace n and h there by $n_{i+1} - n_i$ and $h_{n_{i+1}}$, respectively. By our assumptions, $h_{n_{i+1}} = o(1)$, $(n_{i+1} - n_i)h_{n_{i+1}}^{2d} = o(1)$ and, $(n_{i+1} - n_i)h_{n_{i+1}}^d \to \infty$, so that Lemma 7 indeed applies. The lower bound for the ith term is of the form

$$c_1\left(n_{i+1}h_{n_{i+1}}^d\right)^{-1/2}\exp\left(-c_2 n_{i+1}h_{n_{i+1}}^d\right), \quad i \text{ large enough}, \tag{24}$$

where c_1, c_2 are positive constants for all $\varepsilon > 0$, $\liminf_{\varepsilon \downarrow 0} c_1 > 0$, and $\liminf_{\varepsilon \downarrow 0} c_2 = 0$. Clearly, (24) is at least equal to

$$c_1(M \log\log n_{i+1})^{-1/2}\exp(-c_2 M \log\log n_{i+1})$$

$$\sim c_1(M \log i)^{-1/2}(ai \log i)^{c_2 M},$$

and no tail sum is finite when $c_2 < 1/M$ (i.e., when ε is small enough). This concludes the proof of (23), (21), and Theorem 3.

Proof of Theorem 4. In what follows, we assume that the conditions of Theorem 4 are satisfied. T will be the compact support of f, M is the bound on K, $K = 0$ outside $S_{0,r}$, and $K \geq cI_{S_{0,r}}$. Let A_n be the set of all h_n with $L(h_n) > a \sup_{h>0} L(h)$. There are some measurability problems regarding the h_n that is actually chosen from A_n. These can be sidestepped in a number of ways. We will assume that the choice process is such that h_n is a random variable. We will use the notation F for the distribution function of f, and F_n for the empirical distribution function defined by $X_1, \ldots, X_n$.

The proof of Theorem 4 is based on a number of important lemmas. These are extracted for the convenience of the reader.

LEMMA 9 (Large Deviation Inequalities for the Poisson Distribution). *If X is a Poisson (λ) random variable, then*

$$P\left(|X - \lambda| \geq \lambda\varepsilon\right) \leq 2\exp\left(-\lambda\varepsilon^2/2(1+\varepsilon)\right), \quad \textit{all } \varepsilon > 0.$$

Proof. See (3.3) and its proof.

LEMMA 10. *Let $\{g_\theta\}$ be a collection of functions: $R^1 \to R^1$ parametrized by θ, and let f be a density with compact support. Then*

$$\sup_\theta \left| \int g_\theta \, dF_n - \int g_\theta \, dF \right| \to 0 \quad \textit{almost surely}$$

under the following conditions on our collection:

(i) $\sup_\theta \sup_x |g_\theta(x)| < \infty$ ($\{g_\theta\}$ *is uniformly bounded*);

(ii) $\{g_\theta\}$ *is uniformly equicontinuous.*

Proof. We let $\varepsilon > 0$ be arbitrary, and partition T into rectangles $R_i, 1 \le i \le N$, with $\sup_i \sup_{x,y \in R_i} \|x - y\| \le \delta$, where $\delta > 0$ is chosen so small that $|g_\theta(x) - g_\theta(y)| < \varepsilon$ for all θ, x, y. Let $x_i \in R_i$, $1 \le i \le N$, be arbitrary points in the rectangles. Now,

$$\begin{aligned}
&\sup_\theta \left| \int g_\theta \, dF_n - \int g_\theta \, dF \right| \\
&\quad \le \sum_{i=1}^N \sup_\theta \left| \int_{R_i} g_\theta \, dF_n - \int_{R_i} g_\theta \, dF \right| \\
&\quad \le \sum_{i=1}^N \sup_\theta \left(\int_{R_i} |g_\theta(x) - g_\theta(x_i)| (dF_n + dF) + g_\theta(x_i) \left| \int_{R_i} dF_n - \int_{R_i} dF \right| \right) \\
&\quad \le \sum_{i=1}^N \left(2\varepsilon + \sup_\theta \sup_x |g_\theta(x)| \sup_R \left| \int_R dF_n - \int_R dF \right| \right),
\end{aligned}$$

where R is a rectangle. But by a d-dimensional version of the Glivenko–Cantelli lemma (see, e.g., Kiefer and Wolfowitz, 1958 or Kiefer, 1961),

$$\sup_R \left| \int_R dF_n - \int_R dF \right| \le 2^d \sup_x |F_n(x) - F(x)| \to 0 \quad \text{almost surely.}$$

This concludes the proof of Lemma 10.

LEMMA 11. *For any constants $0 < c_1 \le c_2 < \infty$,*

$$\inf_{c_1 \le h \le c_2} \inf_{x \in T} f * K_h(x) > 0,$$

and

$$\sup_{c_1 \le h \le c_2} \sup_{x \in T} f * K_h(x) \le \frac{M}{c_1^d} < \infty.$$

Proof. The second part of Lemma 11 is obvious. For the first part, we note that

$$f * K_h(x) \ge c \int_{S_{x,rh}} \frac{f}{\lambda(S_{x,rh})} \ge c \int_{S_{x,rc_1}} \frac{f}{(rc_2)^d \lambda(S_{0,1})},$$

where λ is Lebesgue measure. Since T is compact, the collection $\{S_{x,rc_1/2}, x \in T\}$ has a finite subcover of T with spheres centered at $x_1, \ldots, x_N$. Therefore,

$$\inf_{x \in T} f * K_h(x) \ge c_3 \inf_{1 \le i \le N} \int_{S_{x_i,rc_1/2}} f > 0,$$

where c_3 is a positive constant. Since the lower bound does not depend upon h, Lemma 11 is proved.

LEMMA 12. *For all densities f and K satisfying the conditions of Theorem 4, we have*

$$\int \frac{f}{f * K_h} \le M_1 + M_2 h^d,$$

where M_1, M_2 are constants depending upon f, K, and d only.

Proof. First, we note that $f * K_h \ge h^{-d} c \int_{S_{x,hr}} f$. It is always possible to find a cover of T with sets of the form $S_{x,hr/2}$ in such a way that at most $M_3 = M_4 + M_5/h^d$ sets are needed, where the M_i's are constants depending upon f and K only. This is because the smallest closed cube covering T can be covered in such a manner. Let the M_3 centers be called $x_1, x_2, \ldots, x_{M_3}$. It is clear that for all $x \in S_{x_i,hr/2}$, we have $\int_{S_{x,hr}} f \ge \int_{S_{x_i,hr/2}} f$. Thus,

$$\int_T \frac{f}{f * K_h} \le \sum_{i=1}^{M_3} \int_{S_{x_i,hr/2}} \frac{h^d f(x)}{c \int_{S_{x,hr}} f} \, dx$$

$$\le \sum_{i=1}^{M_3} \frac{h^d}{c} \frac{\int_{S_{x_i,hr/2}} f}{\int_{S_{x_i,hr/2}} f} = \frac{M_3 h^d}{c},$$

which was to be shown.

LEMMA 13 (Properties of the $L \log L$ Norm). *Let K and f be densities satisfying the conditions of Theorem 4. Then the following is true:*

(i) $\int f \log f * K_h < \int f \log f$, *all* $h > 0$;

(ii) $\lim_{h \downarrow 0} \int f \log f * K_h = \int f \log f$;

(iii) $\int f \log f * K_h$ *is continuous in* h *on* $(0, \infty)$.

Proof. We note first that $\int f \log f * K_h < \infty$ for all $h > 0$, but that it is possible to have $\int f \log f = \infty$. We start with Jensen's inequality:

$$\int_T f \log\left(\frac{f * K_h}{f}\right) \leq \log\left(\int_T f \cdot \frac{f * K_h}{f}\right) = \log \int_T f * K_h < 0$$

for all $h > 0$.

The proof of (ii) is in two parts. Let $\log_+$ and $\log_-$ denote the positive and negative parts of the log function, respectively. First, by Fatou's lemma and Theorem 2.3,

$$\liminf_{h \downarrow 0} \int f \log_+ f * K_h \geq \int f \liminf_{h \downarrow 0} \log_+ f * K_h = \int f \log_+ f.$$

Next, we need a fact from analysis: for any real number $u \in (0, 1)$, we have

$$\left| \log u + \sum_{j=0}^{J} \frac{(1-u)^j}{j} \right| = \sum_{j=J+1}^{\infty} \frac{(1-u)^j}{j}$$

$$\leq \frac{1}{J+1} \cdot \frac{1}{u}.$$

From this tail estimate and Lemma 12, we conclude that for all integers J,

$$\left| \int f \log_- f * K_h + \int f \sum_{j=0}^{J} \frac{(1 - f * K_h)_+^j}{j} \right|$$

$$\leq \int \frac{f}{f * K_h} \cdot \frac{1}{J+1} \leq \frac{M_1 + M_2 h^d}{J+1},$$

and

$$\left| \int f \log_- f + \int f \sum_{j=0}^{J} \frac{(1-f)_+^j}{j} \right| \leq \frac{\lambda(T)}{J+1},$$

where λ is Lebesgue measure. However, for all integers j, we have $\int f(1 - f * K_h)_+^j \to \int f(1 - f)_+^j$ (by the Lebesgue dominated convergence theorem and Theorem 2.3). All these facts taken together imply that $\int f \log_- f * K_h \to \int f \log_- f$. Thus, since $\log = \log_+ + \log_-$, we have $\liminf_{h \downarrow 0} \int f \log(f * K_h) \geq \int f \log f$, which together with part (i) of this lemma implies (ii).

For part (iii), we consider arbitrary $h, h' > 0$ and start with the following inequality:

$$\left| \int f \log(f * K_h) - \int f \log(f * K_{h'}) \right| \leq \int f \left| \log\left(\frac{f * K_h}{f * K_{h'}} \right) \right|. \tag{25}$$

If $h' > 0$ is fixed and $h \in [h'/2, 2h']$, then $\sup_{x \in T} |\log(f * K_h / f * K_{h'})|$ is uniformly bounded in h by Lemma 11. Let us call this bound c_0. Now, because for $u, v > 0$, $|\log u - \log v| \leq |u - v| / \min(u, v)$, we have the following upper bound for the integrand in (25):

$$c_1 f |f * K_h - f * K_{h'}|. \tag{26}$$

Here c_1 is the bound for $\sup_{h'/2 \leq h \leq 2h'} \sup_{x \in T} 1/(f * K_h)$. The integral of (26) in turn can be bounded as follows:

$$c_1 c_2 \int_{f > c_3} f + c_1 c_3 \int |f * K_h - f * K_{h'}|, \quad \text{all } c_3 > 0, \tag{27}$$

where $c_2 = \sup_{h'/2 \leq h \leq 2h'} \sup_{x \in T} f * K_h$ (this too is finite by Lemma 11). The last term in (27) is $o(1)$ by an argument as in (2.4) of Theorem 2.4 when $h \to h'$. The first term of (27) can be made arbitrarily small by choice of c_3. This concludes the proof of (iii) and of Lemma 13.

LEMMA 14. *Let C be the interval $[c_1, c_2] \subseteq (0, \infty)$, and assume that the conditions of Theorem 4 are satisfied. Then*

$$\sup_{h \in C} \left| \frac{1}{n} \log L(h) - \int f \log(f * K_h) \right| \to 0 \quad \textit{almost surely}.$$

Proof. First, we show that

$$\sup_{h \in C} \left| \frac{1}{n} \log L(h) - \frac{1}{n} \sum_{i=1}^{n} \log(f * K_h(X_i)) \right| \to 0 \quad \text{almost surely}.$$

Indeed, note first that

$$\sup_{h\in C}\sup_{i}|f_{ni}(X_i)-f_n(X_i)| \le \sup_{h\in C}\sup_{x\in T}\sup_{i}|f_{ni}(x)-f_n(x)|$$

$$\le \frac{1}{n}\sup_{h\in C}\sup_{x,i} f_{ni}(x) + \sup_{h\in C}\frac{M}{nh^d} = o(1),$$

and that

$$\sup_{h\in C}\sup_{x}|f_n(x)-f*K_h(x)| \to 0 \quad \text{almost surely}$$

when K is bounded, has compact support, and is a.e. continuous (see, e.g., Devroye and Wagner, 1980 or Bertrand-Retali, 1978). By Lemma 11 about the uniform lower and upper bounds for $f*K_h$, we can conclude that

$$\sup_{h\in C}\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f_{ni}(X_i)}{f*K_h(X_i)}\right) \to 0 \quad \text{almost surely.}$$

Lemma 14 is proved if we can show that

$$\sup_{h\in C}\left|\frac{1}{n}\sum_{i=1}^{n}\log(f*K_h(X_i)) - \int f\log(f*K_h)\right| \to 0 \quad \text{almost surely.}$$

Let us define $g_h = \log(f*K_h)$. We must show that $\sup_{h\in C}|\int g_h\,dF_n - \int g_h\,dF| \to 0$ almost surely. This will be done by verifying the conditions of Lemma 10. First, $\{g_h, h\in C\}$ is uniformly bounded in view of Lemma 11. Thus, we need only check the uniform equicontinuity. But again by Lemma 11, it suffices to verify the equicontinuity of the functions $f*K_h$. Let us take $x, y \in T$. Then,

$$\sup_{h\in C}|f*K_h(x)-f*K_h(y)| \le \sup_{h\in C}\int K_h(z)|f(x-z)-f(y-z)|\,dz$$

$$\le \left(M/c_1^d\right)\int|f(x-z)-f(y-z)|\,dz. \tag{28}$$

But by approximating f in L_1 by a uniformly continuous function f^* with compact support, we can show that this is $o(1)$ as $y \to x$. Because the integral in (28) is thus continuous in y for each fixed x, and because

$x, y \in T$ compact, it must be uniformly continuous in x and y. This concludes the proof of Lemma 14.

LEMMA 15. *Under the conditions of Theorem* 4, *we have* $h_n \to 0$ *almost surely.* (*This is the first half of Theorem* 4.)

Proof. A_n is almost surely not empty because $L(h) = 0$ for all h small enough (this follows from the compactness of the support of K), and $L(h) > 0$ for all h large enough (this follows from $K \geq cI_{S_{0,r}}$).

To prove the Lemma, it suffices to establish that for every $\varepsilon > 0$ there exists a $\delta \in (0, \varepsilon)$ such that

$$\liminf_{n\to\infty}\left(aL(\delta) - \sup_{h\geq\varepsilon} L(h)\right) > 0 \quad \text{almost surely.} \tag{29}$$

(Because if $\gamma \in [\varepsilon, \infty)$, we have $L(\gamma) \leq \sup_{h\geq\varepsilon} L(h) < aL(\delta) \leq a \sup_{h>0} L(h)$, all n large enough, almost surely, and thus $\gamma \notin A_n$.) Now, (29) is satisfied if

$$\limsup_{n\to\infty} \sup_{h\geq\varepsilon} \frac{1}{n} \log L(h) < \liminf_{n\to\infty} \frac{1}{n} \log(aL(\delta)) = \liminf_{n\to\infty} \frac{1}{n} \log L(\delta). \tag{30}$$

By Lemma 14, the right-hand side of (30) is almost surely equal to $\int f \log(f * K_\delta)$. If M_1 is a large positive number, we also have

$$\limsup_{n\to\infty} \sup_{h\geq\varepsilon} \frac{1}{n} \log L(h)$$

$$\leq \max\left(\limsup_{n\to\infty} \sup_{\varepsilon\leq h\leq M_1} \frac{1}{n} \log L(h), \limsup_{n\to\infty} \sup_{h\geq M_1} \log\left(\frac{M}{h^d}\right)\right)$$

$$\leq \max\left(\sup_{\varepsilon\leq h\leq M_1} \int f \log(f * K_h), \log(M/M_1^d)\right)$$

almost surely (by Lemma 14)

$$< \int f \log(f * K_\delta) \quad \text{for some } 0 < \delta < \varepsilon.$$

The last inequality is a consequence of Lemma 13 and our ability to choose M_1 as large as we please. This concludes the proof of Lemma 15.

LEMMA 16. *Under the conditions of Theorem* 4, *we have* $\liminf_{n\to\infty}(nh_n^d/\log n) > 0$ *almost surely. (This is the second half of Theorem* 4.)

Proof. For any $a > 0$, we have

$$P\Big(\sup_{h\in A_n} h < a\Big) \le P\Big(\max_i \min_{j\neq i} \|X_i - X_j\| < \tilde{r}a\Big). \tag{31}$$

This follows from the observation that for $h < (1/\tilde{r})\max_i \min_{j\neq i}\|X_i - X_j\|$, we have $L(h) = 0$, and thus certainly $h \notin A_n$ (because $L(h) > 0$ for all h large enough). Inequality (31) is the starting point. In order to avoid putting too many conditions on f a careful argument is needed. We assume that the support T of f is contained in a closed square Q, which we can take equal to $[0,1]^d$ without loss of generality. We will use (31) with $a = a_n = ((\varepsilon\log n)/n)^{1/d}$ for some small $\varepsilon > 0$ to be picked later. Partition each side of Q into $\underline{1/(\tilde{r}a)}$ intervals of equal length (thus, each interval has length at least equal to ra, and the length $\sim ra$ as $n\to\infty$). The grid of squares has m_n cells B_i, which we assume, again without loss of generality, to be a multiple of 3. Each cell has d coordinates, and each coordinate is an integer between 1 and $\underline{1/(\tilde{r}a)}$, the rank of the interval for that coordinate projection. Let C_i' be the cells with all coordinates of the form $2 + 3j$, $j = 0, 1, 2, \ldots$, and let C_i be all the cells having at least one vertex in common with C_i'. Thus, C_i is a supercell consisting of 3^d original cells, and there are exactly $m_n/3^d$ such cells C_i. In our linear numbering of cells, we will assume that the first $m_n/3^d$ integers give us the indices of the C_i' type cells. Let $p_i = \int_{C_i} f$, $p_i' = \int_{C_i'} f$.

We proced by Poissonization of the sample size. This is entirely a subjective choice. For example, in R^1, the argument could easily be done without Poissonization because the properties of spacings are well understood on the real line. Let N_0, N_1 be independent Poisson random variables with parameters $n - b_n$ and $2b_n$, respectively, where $b_n = \sqrt{M^* n\log n}$, and M^* is a large number to be chosen later. The random variable $N = N_0 + N_1$ is thus Poisson $(n + b_n)$. Given N, draw independent random vectors

$X_1, \ldots, X_N$ from the density f. Now, the right-hand side of (31) is equal to

$$P\left(\bigcap_{i=1}^{n}\left[S_{X_i,ra} \text{ contains at least one } X_j,\ j \neq i,\ j \leq n\right]\right)$$

$$\leq P\left(\bigcap_{i=1}^{m_n/3^d}\left[C_i' \text{ is } \varnothing\right] \cup \left[C_i' \text{ is not } \varnothing, C_i \text{ has at least two } X_j\text{'s}, j \leq n\right]\right)$$

$$\leq P(N < n) + P(N_0 > n)$$

$$+P\left(\bigcap_i [C_i' \text{ has no } X_j, j \leq N_0] \cup [C_i' \text{ has at least one } X_j \text{ with } j \leq N, C_i \text{ has at least two points with index } j \leq N]; N_0 \leq n; N \geq n\right).$$

(32)

By Lemma 9, the first two terms on the right-hand side of (32) do not exceed

$$2\exp\left(-\frac{b_n^2}{2(n+2b_n)}\right) + 2\exp\left(-\frac{b_n^2}{2n}\right) = \frac{4}{n^{M^*/(2+o(1))}},$$

and this is summable in n for $M^* > 2$. Having fixed M^*, we will now bound the last term in (32) by a function summable in n. By (31) and the Borel–Cantelli lemma, this implies that $\sup_{h \in A_n} h < ((\varepsilon \log n)/n)^{1/d}$ finitely often almost surely, which was to be shown.

From the last term of (32), we can drop $N_0 \leq n$, $N \geq n$, thus making it larger. But then, we are left with the probability of the intersection of independent events:

$$\prod_{i=1}^{m_n/3^d}\left(e^{-(n-b_n)p_i'} + \left(1 - e^{-(n+b_n)p_i'} - (n+b_n)p_i' e^{-(n+b_n)p_i' - (n+b_n)(p_i - p_i')}\right)\right)$$

$$\leq \exp\left(\sum_{i=1}^{m_n/3^d}\left(e^{-(n-b_n)p_i'} - e^{-(n+b_n)p_i'}\right) - \sum_{i=1}^{m_n/3^d}(n+b_n)p_i' e^{-(n+b_n)p_i'}\right),$$

(33)

where we have used the inequality $1 + u \leq e^u$ valid for all u. For $0 \leq u \leq v$, we have $e^{-u} - e^{-v} \leq e^{-u}(v-u) \leq v - u$. Thus, the first sum in the exponent of (33) is at most

$$2b_n \sum_{i=1}^{m_n/3^d} p_i' \leq 2b_n.$$

Now, define the following histogram density approximations:

$$g_n(x) = \frac{p_i}{(3\tilde{r}a)^d}; \quad g_n'(x) = \frac{p_i'}{(\tilde{r}a)^d}, \qquad x \in C_i.$$

For the last sum in the exponent of (33), we have

$$\begin{aligned}
&\sum_{i=1}^{m_n/3^d} (n + b_n) p_i' e^{-(n+b_n)p_i} \\
&\quad \sim 3^{-d} \sum_{i=1}^{m_n/3^d} \int_{C_i} (n + b_n) g_n'(x) e^{-(n+b_n)(3\tilde{r}a)^d g_n(x)} dx \\
&\quad \sim 3^{-d} \int_Q n g_n'(x) e^{-(n+b_n)(3\tilde{r}a)^d g_n(x)} dx. \qquad (34)
\end{aligned}$$

But by Fatou's lemma and Theorem 2.2,

$$\liminf_{n\to\infty} \frac{(34)}{2b_n} \geq \frac{1}{2} 3^{-d} \int_Q f(x) \liminf_{n\to\infty} \sqrt{\frac{n}{M^* \log n}}\, e^{-(n+b_n)(3\tilde{r}a)^d g_n(x)}$$

which is infinite if on a set of nonzero f-measure,

$$\liminf_{n\to\infty} \sqrt{\frac{n}{\log n}} \exp\bigl(-(n + b_n)(3\tilde{r}a)^d g_n(x)\bigr) = \infty. \qquad (35)$$

By Theorem 2.2, $g_n \to f$ for almost all x. Also, $(n + b_n)(3\tilde{r}a)^d \sim (3\tilde{r})^d \varepsilon \log n$ and the term in (35) is

$$\sqrt{\frac{n}{\log n}} \exp\bigl(-(3\tilde{r})^d f(x)(\varepsilon + o(1)) \log n\bigr), \quad \text{almost all } x.$$

This tends to ∞ on a set of positive f-measure if ε is chosen small enough. Thus, for all n large enough, (33) is at most equal to $\exp(-\sqrt{n \log n})$, and this is summable in n. This concludes the proof of Lemma 16.

6. INVARIANT DENSITY ESTIMATION

Invariance of density estimates under certain transformations is an issue first suggested in the work of Wertz (1974a, 1976). In this section, we will explain why invariance motivates us to use automatic density estimates in general.

Let ϕ be a real-valued function on R, monotonically increasing, one-to-one and onto, and let ϕ and its inverse be absolutely continuous on finite intervals. We say that a density estimate f_n on R is *ϕ-invariant* if for all n, and all $x, x_1, \ldots, x_n \in R^{(n+1)}$,

$$f_n(\phi(x), \phi(x_1), \ldots, \phi(x_n)) = (\phi^{-1}(x))' f_n(x, x_1, \ldots, x_n).$$

This can be rephrased as follows. Assume that we have constructed a density estimate $f_n(x, X_1, \ldots, X_n)$. Then, an estimate of the density of $Y = \psi(X_1)$ can be obtained in two ways:

(i) By an ordinary transformation of densities: this gives, if $\phi = \psi^{-1}$,

$$\phi'(y) f_n(\phi(y), \phi(y_1), \ldots, \phi(y_n)), \qquad y, y_1, \ldots, y_n \in R^{(n+1)}.$$

(ii) By constructing a new estimate based on $\psi(X_1), \ldots, \psi(X_n)$:

$$f_n(\psi(x), \psi(x_1), \ldots, \psi(x_n)) = f_n(y, y_1, \ldots, y_n),$$

$$y, y_1, \ldots, y_n \in R^{(n+1)}.$$

Essentially, ϕ-invariance means that both estimates are identical. To put it in yet another way, ϕ-invariance means that for all Borel sets B,

$$\int_B f_n(y, y_1, \ldots, y_n)\, dy = \int_{\phi[B]} f_n(y, \phi(y_1), \ldots, \phi(y_n))\, dy,$$

where $\phi[B]$ is the set of all values $\phi(y)$, $y \in B$.

In particular, we are interested in *translation invariance* ($\phi(x) = x + a, a \in R$), and *scale invariance* ($\phi(x) = bx, b > 0$). Translation invariance is nearly always taken for granted, and estimates that are not translation invariant seem somehow peculiar. Translation invariance is basically equivalent to the requirement that f_n be a function of $x - X_1, \ldots, x - X_n$ only. Obviously, all kernel estimates are translation invariant, as long as h is fixed. The same remains true if we allow h to depend upon all pairwise differences $X_i - X_j = (x - X_j) - (x - X_i)$. Unfortunately, unless we modify its definition, the histogram estimate is not translation invariant. *Dirac delta function estimates* (Walter and Blum, 1979) are estimates of the form

$$f_n(x) = \sum_{i=1}^{n} w_{ni} K_{ni}(x, X_i),$$

where the w_{ni}'s are weights, and the K_{ni}'s are given real-valued functions. Estimates of this form are translation invariant only if $K_{ni}(x, y) = K_{ni}^*(x - y)$ for some functions K_{ni}^*. This is why only the latter form is treated in this book. In Chapter 12, we will see that orthogonal series estimates on the real line are not translation invariant. We should also note that without translation invariance, it is difficult to enforce the condition $\int f_n = 1$.

Scale invariant density estimates satisfy

$$bf_n(bx, bX_1, \ldots, bX_n) = f_n(x, X_1, \ldots, X_n), \quad \text{all } b > 0.$$

Unfortunately, the standard kernel estimate with fixed h is not scale invariant for all kernels K because the condition

$$\frac{1}{b}\left(\frac{1}{h}K\left(\frac{x-y}{h}\right)\right) = \frac{1}{h}K\left(\frac{bx-by}{h}\right), \quad \text{all } x, y;\ R, b, h > 0,$$

is only fulfilled if $K(x) = c/x$, or $K(x) = c/|x|$ for some $c \in R$. These choices of kernels lead to kernel estimates with horrible properties: for example, for $K(x) = c/|x|$, we have $\int |f_n| = \infty$, all n, and $E(f_n(x)) = \infty$ for almost all x for which $f(x) > 0$. The standard conditions on the kernel K, that is, $\int |K| < \infty$, $\int K = 1$, cannot lead to kernel estimates that are scale invariant unless h is data-dependent. In particular, we obtain scale invariance if h depends upon the data in such a way that

$$h(bX_1, \ldots, bX_n) = bh(X_1, \ldots, X_n).$$

This is one of the main arguments for studying automatic kernel estimates. Such functions h include for example,

$$c\left(\sum_{i,j} |X_i - X_j|^p\right)^{1/p}, \qquad p > 0,$$

$$c\left(X_{(b_n)} - X_{(a_n)}\right), \quad \text{where } 1 \le a_n \le b_n \le n,\ X_{(i)} \text{ is the } i\text{th order statistic of } X_1, \ldots, X_n.$$

In both cases, c is a suitably chosen function of n only.

Invariance with respect to other transformations is rarely needed: there are applications in which density estimates may be required on a linear and a logarithmic scale, and it would be quite distressing to observe that the standard log-transform of the linear scale estimate does not correspond to the estimate constructed with the logarithmically transformed data. If there is inequality, which estimate should one choose? Another question we have not tackled here is that of the existence of meaningful (consistent, etc.) ϕ-invariant estimates for a given ϕ.

7. RATE OF CONVERGENCE FOR AUTOMATIC KERNEL ESTIMATES

In this section we give a general theorem that allows us to infer things about the rate of convergence of an automatic kernel estimate provided that we know something about the asymptotic behavior of h, in particular, its closeness to a deterministic sequence a_n. We impose a new condition on the kernel, but emphasize that this is merely for technical convenience.

THEOREM 5. *Assume that f_n is an automatic kernel estimate with smoothing factor h and kernel K, where $\int K = 1$, and*

$$\int |K_u - K| \le C(1 - u^{-d}), \qquad u \ge 1,$$

for some constant C. If f_{na} is the standard kernel estimate with smoothing factor a and the same kernel K, then

$$\left| \int |f_n - f| - \int |f_{na} - f| \right| \le C\left(1 - \min^d\left(\frac{a}{h}, \frac{h}{a}\right)\right).$$

Proof. Without loss of generality, assume that $a \le h$. Then

$$\begin{aligned}
&\left| \int |f_n - f| - \int |f_{na} - f| \right| \\
&\le \int |f_n - f_{na}| \\
&\le \frac{1}{n} \sum_{i=1}^{n} \int \left| h^{-d} K\left(\frac{x - X_i}{h}\right) - a^{-d} K\left(\frac{x - X_i}{a}\right) \right| dx \\
&= \int \left| \left(\frac{a}{h}\right)^d K\left(\frac{a}{h} x\right) - K(x) \right| dx \\
&\le C\left(1 - \left(\frac{a}{h}\right)^d\right).
\end{aligned}$$

REMARK. The condition put on K is satisfied for most densities K, in particular for all densities K on R^d that are nonincreasing along rays, that is, $K(ux) \le K(x)$, all $x \in R^d, u \ge 1$. In the latter case, C can be taken

equal to 2, as can be seen from the following simple argument:

$$\int |K_u - K| \le \int \left| u^{-d} K\left(\frac{x}{u}\right) - u^{-d} K(x) \right| dx + \int |u^{-d} K(x) - K(x)| dx$$

$$= u^{-d} \int \left(K\left(\frac{x}{u}\right) - K(x) \right) dx + (1 - u^{-d})$$

$$= 2(1 - u^{-d}), u \ge 1.$$

The inequality of Theorem 5 does not impose any conditions on the smoothing factor, such as convergence to zero, and so on. It has many uses, and we will list only a few in some Lemmas.

LEMMA 17. *For any automatic kernel estimate f_n, with kernel K as in Theorem* 5,

$$E\left(\int |f_n - f| \right) \le E\left(\int |f_{na_n} - f| \right) + C\varepsilon_n + 2P\left(1 - \min^d\left(\frac{a_n}{h}, \frac{h}{a_n} \right) \ge \varepsilon_n \right),$$

where a_n, ε_n are positive number sequences, and f_{na_n} is the standard kernel estimate with smoothing factor a_n.

This Lemma is extremely useful. For example, consider smoothing factors of the form

$$h = a_n \frac{2}{n} \sum_{i=1}^{n/2} |X_{2i-1} - X_{2i}|.$$

Here a_n is an arbitrary sequences of positive numbers. We could have suggested a double sum over all $|X_i - X_j|$, but this could be computationally unfeasible, and would in any case cloud the issue at stake. We have the following lemma.

LEMMA 18. *If an automatic kernel estimate is constructed with nonnegative kernel, vanishing outside $[-1, 1]$, and satisfying the condition of Theorem* 5, *and if h is chosen as in the previous paragraph for some arbitrary sequence of positive numbers a_n, and if $E(|X_1|^p) < \infty$ for some $p > 4$, then*

$$E\left(\int |f_n - f| \right) \sim E\left(\int |f_{n\, a_n E(|X_1 - X_2|)} - f| \right)$$

in the notation of Lemma 17. *There are no other conditions on f apart from the given moment condition.*

Proof. We will apply Lemma 17 directly. By Theorem 5.2, the result follows if we can find a sequence ε_n such that $\varepsilon_n = o(n^{-2/5})$ and $P(|h/(a_n E(|X_1 - X_2|)) - 1| > \varepsilon_n) = o(n^{-2/5})$. By Chebyshev's inequality, the latter probability is not greater than

$$E\left(\varepsilon_n^{-p}\left|\frac{2}{n}\sum_{i=1}^{n/2}(Y_i - E(Y_i))\right|^p\right) \quad (\text{where } Y_i = |X_{2i-1} - X_{2i}|)$$

$$\leq C_p \varepsilon_n^{-p}\left(\frac{2}{n}\right)^p E\left(\left(\sum_{i=1}^{n/2}(Y_i - E(Y_i))^2\right)^{p/2}\right)$$

(valid $p \geq 2$, some $C_p > 0$ by the inequality of Marcinkiewicz and Zygmund, see Lemma 5.27)

$$\leq C_p \varepsilon_n^{-p}\left(\frac{2}{n}\right)^p \left(\frac{n}{2}\right)^{p/2} 1^{n/2} \sum_{i=1} E(|Y_i - E(Y_i)|^p) \quad (\text{by the } c_r\text{-inequality})$$

$$\leq C_p \varepsilon_n^{-p}\left(\frac{2}{n}\right)^{p/2} 2^{p-1} E(|Y_1|^p)$$

$$\leq A\varepsilon_n^{-p} n^{-p/2},$$

where A is a finite constant. If we pick $\varepsilon_n \sim n^{-p/2(p+1)}$, then $\varepsilon_n = o(n^{-2/5})$ for $p > 4$, and $\varepsilon_n^{-p} n^{-p/2}$ decreases as ε_n times a constant. This concludes the proof of Lemma 18.

REMARK. From Theorem 5, Theorem 3.1 and an argument similar to that used to obtain Lemma 17, we see that if there exists a sequence a_n with $a_n \to 0$, $na_n^d \to \infty$, such that $h/a_n \to 1$ in probability (almost surely), then $\int |f_n - f| \to 0$ in probability (almost surely), provided that the kernel integrates to 1, is absolutely integrable, and satisfies the conditions of Theorem 5. Theorem 6.1 is more general, because no extra conditions are imposed on K, and the existence of a centering sequence a_n is not assumed (this allows for more variability in the random variable h).

REFERENCES

G. Bennett (1962). Probability inequalities for the sum of independent random variables, *Journal of the American Statistical Association* **57**, pp. 33–45.

M. Bertrand-Retali (1978). Convergence uniforme d'un estimateur de la densité par la méthode du noyau, *Revue Roumaine de Mathématiques Pures et Appliquées* **23**, pp. 361–385.

A. W. Bowman (1982). A comparative study of some kernel-based nonparametric density estimators, Manchester-Sheffield School of Probability and Statistics, Research Report No. 84/AWB/1.

C. Bretagnolle and C. Huber (1979). Estimation des densités: risque minimax, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **47**, pp. 119–137.

Y. S. Chow, S. Geman, and L. D. Wu (1983). Consistent cross-validated density estimation, *Annals of Statistics* **11**, pp. 25–38.

P. Deheuvels (1974). Conditions nécessaires et suffisantes de convergence ponctuelle presque sûre et uniforme presque sûre des estimateurs de la densité, *Comptes Rendus Académie des Sciences de Paris Série A* **178**, pp. 1217–1220.

P. Deheuvels (1977). Estimation non paramétrique de la densité par histogrammes généralisés, *Revue de Statistique Appliquée* **25**, pp. 5–42.

P. Deheuvels and P. Hominal (1980). Estimation automatique de la densite, *Revue de Statistique Appliquée* **28**, pp. 25–55.

L. Devroye and T. J. Wagner (1980). The strong uniform consistency of kernel density estimates, in *Multivariate Analysis V*, P. R. Krishnaiah (Ed.), North-Holland, New York, pp. 59–77.

R. P. W. Duin (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions, *IEEE Transactions on Computers* **C-25**, pp. 1175–1179.

S. Geman (1981). Sieves for nonparametric estimation of densities and regressions, Reports in Pattern Analysis No. 99, Division of Applied Mathematics, Brown University, Providence, Rhode Island.

S. Geman and C.-R. Hwang (1982). Nonparametric maximum likelihood estimation by the method of sieves, *Annals of Statistics* **10**, pp. 401–414.

J. D. F. Habbema, J. Hermans, and K. Vandenbrock (1974). A stepwise discriminant analysis program using density estimation in *COMPSTAT 1974*, G. Bruckmann (Ed.), Physica Verlag, Wien, pp. 101–110.

P. Hall (1982a). Cross-validation in density estimation, *Biometrika* **69**, pp. 383–390.

P. Hall (1982b). Limit theorems for stochastic measures of the accuracy of nonparametric density estimators, *Stochastic Processes and Applications* **13**, pp. 11–25.

P. Hall (1983a). Large-sample optimality of least squares cross-validation in density estimation, *Annals of Statistics* **11**, pp. 1156–1174.

P. Hall (1983b). Asymptotic theory of minimum integrated square error for multivariate density estimation, Proceedings of the Sixth International Symposium on Multivariate Analysis, Pittsburgh.

J. Kiefer (1961). On large deviations of the empiric d.f. of vector chance variables and a law of the iterated logarithm, *Pacific Journal of Mathematics* **11**, pp. 649–660.

J. Kiefer and J. Wolfowitz (1958). On the deviations of the empiric distribution function of vector chance variables, *Transactions of the American Mathematical Society* **87**, pp. 173–186.

E. A. Nadaraya (1974). On the integral mean square error of some nonparametric estimates for the density function, *Theory of Probability and Its Applications* **19**, pp. 133–141.

M. Rosenblatt (1956). Remarks on some nonparametric estimates of a density function, *Annals of Mathematical Statistics* **27**, pp. 832–837.

M. Rosenblatt (1971). Curve estimates, *Annals of Mathematical Statistics* **42**, pp. 1815–1842.

M. Rudemo (1982). Empirical choice of histogram and kernel density estimators, *Scandinavian Journal of Statistics* **9**, pp. 65–78.

E. F. Schuster and G. G. Gregory (1978). Choosing the shape factor(s) when estimating a density, Technical Report, Department of Mathematics, University of Texas, El Paso, Texas.

E. F. Schuster and G. G. Gregory (1981). On the nonconsistency of maximum likelihood nonparametric density estimators, in *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, W. F. Eddy (Ed.), Springer-Verlag, New York, pp. 295–298.

D. W. Scott (1979). Optimal data-based histograms, *Biometrika* **66**, pp. 605–610.

D. W. Scott, R. A. Tapia, and J. R. Thompson (1977). Kernel density estimation revisited, *Journal of Nonlinear Analysis, Theory, Methods and Applications* **1**, pp. 339–372.

D. W. Scott and L. E. Factor (1981). Monte Carlo study of three data-based nonparametric probability density estimators, *Journal of the American Statistical Association* **76**, pp. 9–15.

E. Seneta (1976). *Regularly Varying Functions*, Lecture Notes in Mathematics #508, Springer-Verlag, Heidelberg.

B. W. Silverman (1978). Choosing the window width when estimating a density, *Biometrika* **65**, pp. 1–11.

C. J. Stone (1983). An asymptotically efficient histogram selection rule, Proceedings of the Neyman–Keifer meeting.

C. J. Stone (1984). An asymptotically optimal window selection rule for kernel density estimates, Technical Report, University of California at Berkeley.

R. A. Tapia and J. R. Thompson (1978). *Nonparametric Probability Density Estimation*, The Johns Hopkins University Press, Baltimore, Maryland.

T. J. Wagner (1975). Nonparametric estimates of proability densities, *IEEE Transactions on Information Theory* **IT-21**, pp. 438–440.

G. Walter and J. Blum (1979). Probability density estimation using delta sequences, *Annals of Statistics* **7**, pp. 328–340.

W. Wertz (1974a). Invariante und Optimale Dichteschätzungen, *Mathematica Balkanica* **4**, pp. 707–722.

W. Wertz (1974b). On the existence of density estimators, *Studia Scientiarium Mathematicarum Hungarica* **9**, pp. 45–50.

W. Wertz (1976). Invariant density estimation, *Monatshefte für Mathematik* **81**, pp. 315–324.

M. Woodroofe (1970). On choosing a delta-sequence, *Annals of Mathematical Statistics* **41**, pp. 1665–1671.

CHAPTER 7

Estimates Related to the Kernel Estimate and the Histogram Estimate

1. INTRODUCTION

The density estimates in this chapter are all valid densities on R^d. They are mainly generalizations of the kernel and histogram estimates with some extra features, such as:

(i) improved small sample performance;

(ii) robustness;

(iii) simple recursive definition;

(iv) locally adapted smoothing.

In general, the consistency of these estimates is easy to establish although equivalence results of the format of Theorem 3.1 are not available at this moment. In Chapters 3–5 we obtained a solid foundation for comparing the kernel and histogram estimates. For the generalizations of these estimates, much less is known about the rates of convergence. While consistency is normally a routine matter, the rate of convergence is not.

The estimates are quite arbitrarily grouped as follows:

1. Variable kernel estimates.
2. Recursive kernel estimates.
3. Maximum likelihood estimates.
4. Varible histogram estimates.
5. Kernel estimates with reduced bias.
6. Grenander's estimate for monotone densities.

2. VARIABLE KERNEL ESTIMATES

In 1977, Breiman, Meisel, and Purcell proposed the following estimate:

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} H_{ni}^{-d} K\left(\frac{x - X_i}{H_{ni}}\right), \tag{1}$$

where K is a given kernel, and H_{ni} is the distance between X_i and its k_nth nearest neighbor among $X_1, \ldots, X_n$. Estimate (1) is a density in x, and seems to have a locally adapted smoothing parameter because, roughly speaking, H_{ni} is large where f is small and vice versa. Of course, one still has to choose k_n, so this estimate could hardly be called automatic.

Estimate (1) probably emerged from the *nearest-neighbor estimate* which could be thought of as (1) with a replacement of H_{ni} by $H_n(x)$, the distance of x to its k_nth nearest neighbor among $X_1, \ldots, X_n$. See, for example, Moore and Yackel (1977), Mack and Rosenblatt (1979), and Loftsgaarden and Quesenberry (1965) (who were the first to define the nearest-neighbor estimate for the special choice $K(x) = I_{S_{0,1}}(x)/\lambda(S_{0,1})$). However, using $H_n(x)$ instead of H_{ni} destroys the density property: in fact, we have $\int f_n = \infty$ for all n.

The idea of a locally adapted smoothing parameter is worthy of further study. Our analysis of Chapter 5 shows that the smoothing factor should be large when $1/f$ and $|f''|$ are small. In areas of high curvature of f (large $|f''|$) or of small values of f, smaller smoothing factors are needed. The quantity H_{ni} does not take the curvature component into consideration and is thus inherently asymptotically suboptimal. In fact, for fixed curvature, the smoothing factor should increase, not decrease, with increasing values of f, as suggested by the nearest-neighbor methods.

Nevertheless, the experimental results reported for (1) in Breiman et al. (1977), Habbema et al. (1978), and Raatgever et al. (1978) are promising, to say the least. Yet what has eluded most researchers is a firm grasp on the properties of (1), which is a sum of dependent random variables. The only consistency result reported in the literature states that $\sup_x |f_n - f| \to 0$ completely when f is uniformly continuous, K is the uniform density on the unit sphere, $k_n/n \to 0$, and $k_n/\log n \to \infty$ (Devroye and Penrod, 1982). Abramson (1982) has suggested another method for choosing H_{ni}, and although his method lets H_{ni} depend upon x, it has an obvious x-independent extension. What is needed here is a consistency theorem in the spirit of Theorem 6.2 (which can be considered as a special case in which $H_{n1} = \cdots = H_{nn} = H_n$ is random), with conditions on, say, the quantiles of the H_{ni} sequence.

Another problem worthy of further investigation is that of the choice of H_{ni}. Nearest-neighbor methods, such as the one used by Breiman et al. (1977), let the smoothing factor increase with f. There is no direct dependence upon the curvature. But as we have seen in Chapter 5, even though the results developed there are global, the smoothing factor should ideally depend upon f and f''.

3. RECURSIVE KERNEL ESTIMATES

Recursively defined estimates offer two advantages: data need not be stored, and the estimates are easy to update when new data become available. In the former case, we are presumably only interested in the value of f at several fixed x's. One can hardly expect simple recursive versions of the Parzen–Rosenblatt kernel estimate to perform as well was the nonrecursive original estimate.

The most frequently mentioned estimate defined by

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} h_i^{-d} K\left(\frac{x - X_i}{h_i}\right) \tag{2}$$

is due to Wolverton and Wagner (1969a, b) and Yamato (1971) (for its theoretical analysis see Davies, 1973; Deheuvels, 1973a, b, 1974; Carroll, 1976; Ahmad and Lin, 1976; Devroye, 1979; Wegman and Davies, 1979; Györfi, 1981). Here, h_n is a sequence of positive smoothing factors.

Deheuvels (1973a, b; 1974) also proposed other generalizations including

$$f_n(x) = \frac{\sum_{i=1}^{n} K((x - X_i)/h_i)}{\sum_{i=1}^{n} h_i^d} \tag{3}$$

and

$$f_n(x) = \frac{\sum_{i=1}^{n} a_i K((x - X_i)/h_i)}{\sum_{i=1}^{n} a_i h_i^d}, \tag{4}$$

where $a_i = g(h_i)$ for some positive-valued function g. In particular, he

showed that the asymptotic variance of (3) is better than that of (2) or (4) (with $a_i \neq 1$).

Deheuvels (1979) derives the optimal asymptotic values for h_i and a_i under a mean integrated square error criterion, and concludes that when K is a symmetric density on R^1, estimate (2) is optimal. Estimates (2) and (3) will therefore be considered in some detail below. We would also like to draw the attention to the general classes of estimates introduced by Banon (1976) and Rejtö and Révész (1973).

For $a_i = 1/\sqrt{h_i^d}$, (4) comes close to an estimate studied in Wegman and Davies (1979). Finally, we would like to point out the work of Isogai (1978, 1979, 1982) who considers estimates defined recursively by

$$f_{n+1}(x) = f_n(x) + a_{n+1}\left(h_{n+1}^{-d} K\left(\frac{x - X_{n+1}}{h_{n+1}}\right) - f_n(x)\right), \tag{5}$$

where $a_1 = 1$, $0 < a_n \leq 1$, $a_n \to 0$ and $\sum_{n=1}^{\infty} a_n = \infty$. He gives sufficient conditions for various types of consistency. Note that for the choice $a_n = 1/n$, (5) defines (2).

THEOREM 1. *Let K be a bounded density with integrable radial majorant (see Theorem 2.3), and let $\{h_n\}$ be a sequence of nonnegative numbers. Let f_n be estimate (3). Then, the following are equivalent:*

A. $f_n \to f$ *almost surely, almost all x, all f;*
B. $f_n \to f$ *in probability, almost all x, some f;*
C. $\sum_{n=1}^{\infty} h_n^d = \infty$ *and* $\lim_{n\to\infty} \sum_{i=1}^{n} h_i^d I_{[h_i > \varepsilon]} / \sum_{i=1}^{n} h_i^d = 0$ *for all* $\varepsilon > 0$;
D. $\int |f_n - f| \to 0$ *almost surely, all f;*
E. $\int |f_n - f| \to 0$ *in probability, some f.*

Theorem 1 will be proved here in full. Parts of the proof are taken from Deheuvels (1973a, b; 1974) and Devroye (1979). We would like to point out that weak and strong pointwise consistency are equivalent, a property not shared by the standard kernel estimate (see Theorem 6.3).

Note that C is implied by the conditions $h_n = o(1)$ and $\sum_{n=1}^{\infty} h_n^d = \infty$, but that it does not imply that $h_n = o(1)$.

For the proof of Theorem 1, we will base ourselves on a few key lemmas.

LEMMA 1. *Any random variable X with absolute moments $\mu_r = E(|X|^r)$ satisfies*

$$\mu_1 \geq \sqrt{\mu_2^3 / \mu_4}.$$

Proof. The function $1/x + ax^2$ on $(0, \infty)$ (for fixed $a > 0$) is minimal for $x^3 = 1/2a$. Thus,

$$\frac{x + ax^4}{x^2} \geq (2a)^{1/3} + a\left(\frac{1}{2a}\right)^{2/3} = \frac{3}{2}(2a)^{1/3}.$$

From this, we have, replacing x by $|X|$, and taking expectations,

$$\mu_1 \geq \tfrac{3}{2}(2a)^{1/3}\mu_2 - a\mu_4.$$

The lower bound considered as a function of a is maximal for $a = \frac{1}{2}(\mu_2/\mu_4)^{3/2}$. Resubstitution into the bound gives

$$\mu_1 \geq \frac{\mu_2^{3/2}}{\mu_4^{1/2}}.$$

LEMMA 2. *If K is a bounded density with integrable radial majorant and $h \downarrow 0$, then*

$$f * K_h^p \to f\int K^p \quad \textit{for almost all } x, \textit{ and all } p > 0.$$

Proof. This is a consequence of Theorem 2.3.

Proof of Theorem 1. By Theorem 2.8, A $\Rightarrow$ B $\Rightarrow$ E and A $\Rightarrow$ D $\Rightarrow$ E. We will first show C $\Rightarrow$ A, and we will conclude later by proving that E $\Rightarrow$ C.

Define

$$f_n^*(x) = \frac{\sum_{i=1}^n K\big((x - X_i)/h_i\big) I_{[h_i \leq \varepsilon]}}{\sum_{i=1}^n h_i^d}$$

for some $\varepsilon > 0$. Then, by C,

$$|f_n - f_n^*| \leq \frac{M \sum_{i=1}^n I_{[h_i > \varepsilon]}}{\sum_{i=1}^n h_i^d} = o(1),$$

where M is the bound for K. Also,

$$|f - f_n^*| \le \left| \frac{\sum_{i=1}^n V_i}{\sum_{i=1}^n h_i^d} \right| + \frac{\left(\sum_{i=1}^n h_i^d |f * K_{h_i} - f| I_{[h_i \le \varepsilon]} + \sum_{i=1}^n h_i^d I_{[h_i > \varepsilon]} f \right)}{\sum_{i=1}^n h_i^d},$$

where

$$V_i = h_i^d \left(h_i^{-d} K\left((x - X_i)/h_i\right) - f * K_{h_i}(x) \right) I_{[h_i \le \varepsilon]}, \qquad 1 \le i \le n,$$

are independent random variables. The last term on the right-hand side of the inequality consists of a term that can be made small by choosing ε small (in view of Lemma 2), and a term that is $o(1)$ for all $\varepsilon > 0$ (by C), and this for almost all x. Thus, A follows if we can show that $\sum_{i=1}^n V_i / \sum_{i=1}^n h_i^d \to 0$ almost surely for almost all x. Such a random variable tends to 0 almost surely when

$$\sum_{n=1}^{\infty} h_n^d = \infty, \quad \sum_{n=1}^{\infty} \frac{E\left(V_n^2\right)}{\left(\sum_{i=1}^n h_i^d \right)^2} < \infty \tag{6}$$

(see, e.g., Loève, 1963, p. 253). But we note that for almost all x,

$$E\left(V_n^2\right) \le h_n^{2d} I_{[h_n \le \varepsilon]} h_n{}^d f * K_{h_n}^2 \le h_n^d I_{[h_n \le \varepsilon]} \left(f \int K^2 + 1 \right)$$

by our choice of ε and Lemma 2. Thus, we need only verify that

$$\sum_{n=1}^{\infty} \frac{h_n^d}{\left(\sum_{i=1}^n h_i^d \right)^2} < \infty.$$

But this is true because, if $h_1 > 0$ (without loss of generality),

$$\sum_{n=1}^{\infty} \frac{h_n^d}{\left(\sum_{i=1}^{n} h_i^d\right)^2} \le h_1^{-d} + \sum_{n=2}^{\infty} \frac{h_n^d}{\left(\sum_{i=1}^{n} h_i^d \sum_{i=1}^{n-1} h_i^d\right)}$$

$$= h_1^{-d} + \sum_{n=2}^{\infty} \left(\left(\sum_{i=1}^{n-1} h_i^d\right)^{-1} - \left(\sum_{i=1}^{n} h_i^d\right)^{-1} \right) = 2h_1^{-d} < \infty.$$

This concludes the proof of C $\Rightarrow$ A.

We will now show that $E \Rightarrow \sum_{n=1}^{\infty} h_n^d = \infty$. We proceed by contradiction. Assume that the sum is finite and equal to s. Thus, we have $h_n \to 0$ and, as established in the first part of this theorem, $|E(f_n) - f| \to 0$ almost all x. Now, by Fatou's lemma and E, $0 = \liminf_{n\to\infty} \int E(|f_n - f|) \ge \int \liminf_{n\to\infty} E(|f_n - f|)$, and thus,

$$\liminf_{n\to\infty} E(|f_n - E(f_n)|) = 0, \quad \text{almost all } x. \tag{7}$$

Assume first that there is an infinite sequence of positive h_n's. In view of Lemma 1, we obtain a contradiction with (7) if we can show that

$$\limsup_{n\to\infty} E(|f_n - E(f_n)|^4) < \infty, \quad \text{almost all } x, \tag{8}$$

and

$$\liminf_{n\to\infty} E(|f_n - E(f_n)|^2) > 0, \quad \text{almost all } x. \tag{9}$$

Let us define $Y_i = K((x - X_i)/h_i)$, $Z_i = Y_i - E(Y_i)$. By Lemma 2, $E(Y_i^r) \sim h_i^d f(x) \int K^r$, all $r > 0$, almost all x. From this, we deduce that $E(Z_i^2) \sim E(Y_i^2) \sim h_i^d f(x) \int K^2$, almost all x, and that $E(Z_i^4) \sim E(Y_i^4) \sim h_i^d f(x) \int K^4$, almost all x. To verify (8), we note that for almost all x, as $n \to \infty$,

$$E(|f_n - E(f_n)|^4)\left(\sum_{i=1}^{n} h_i^d\right)^4 \sim s^4 E\left(\left(\sum_{i=1}^{n} Z_i\right)^4\right)$$

$$\le s^4 E\left(\sum_{i=1}^{n} Z_i^4 + 6\sum_{i=1}^{n}\sum_{j=1}^{n} Z_i^2 Z_j^2\right) \le cs^8$$

for some constant c depending upon x. Here we used the fact that $E(Z_i^2)/h_i^d$ and $E(Z_i^4)/h_i^d$ are uniformly bounded in i for almost all x.

Also, $E(Z_i^2) \geq (f(x)h_i^d \int K^2)/2$ for all $i \geq N(x)$, almost all x. Thus,

$$E\left(|f_n - E(f_n)|^2\right) \geq s^{-2} \sum_{i=N(x)}^{n} E\left(Z_i^2\right) \geq \frac{f(x)\int K^2}{2s^2} \sum_{i=N(x)}^{n} h_i^d$$

$$\to \frac{f(x)K^2}{2s^2} \sum_{i=N(x)}^{\infty} h_i^d > 0, \quad \text{almost all } x.$$

This shows (9). If $h_n = 0$ for all $n \geq N$, then a contradiction with (7) is also easily obtained. Thus, the first condition of C must hold.

For the proof of the second half of condition C, we use characteristic functions. Let ϕ and ψ be the characteristic functions of f and K. Since $E(\int |f_n - f|) \to 0$, and $E(\int |f_n - f|) \geq \int |E(f_n) - f|$, we conclude that the characteristic function of $E(f_n)$ must tend to the characteristic function of f. The characteristic function of $E(f_n)$ is

$$\phi_n(t) = \frac{\sum_{i=1}^{n} h_i^d \phi(t)\psi(h_i t)}{\sum_{i=1}^{n} h_i^d}, \qquad t \in R^d.$$

We know that $\sum_{n=1}^{\infty} h_n^d = \infty$ (because it is implied by E).

Assume that there exist positive numbers a and b such that along a subsequence of n, $\sum_{i=1}^{n} h_i^d I_{[h_i > a]} / \sum_{i=1}^{n} h_i^d \geq b$. Because

$$\phi_n(t) - \phi(t) = \phi(t) \frac{\sum_{i=1}^{n} h_i^d (\psi(h_i t) - 1)}{\sum_{i=1}^{n} h_i^d} \to 0,$$

we have for all t small enough,

$$\frac{\sum_{i=1}^{n} h_i^d (\psi(h_i t) - 1)}{\sum_{i=1}^{n} h_i^d} \to 0.$$

But the real part of this expression is, for the n in our subsequence, at most equal to

$$\sup_{h \geq a}(\mathrm{Re}(\psi(ht)-1))\frac{\sum_{i=1}^{n} h_i^d I_{[h_i \geq a]}}{\sum_{i=1}^{n} h_i^d} \leq -b \sup_{h \geq a}(\mathrm{Re}(\psi(ht)-1)).$$

Thus, we must have $\sup_{h \geq a}\mathrm{Re}(\psi(ht)) = 1$ for all t small enough. By the continuity of ψ, this implies that $\mathrm{Re}(\psi(ht)) = 1$ for some $h \geq a$, $t \neq 0$. But this is impossible, because ψ is the characteristic function of a random variable with a density. Thus, we have a contradiction, and therefore E $\Rightarrow$ C.

For estimate (2) we again have different strong pointwise and strong L_1 behavior. Although all types of L_1 convergence seem equivalent, this will not be shown here. We briefly state a theorem with some sufficient conditions of convergence. The necessity of these conditions was shown by Deheuvels (1974) under various regularity conditions on f, K, and h_n. The necessity can be established in all generality by using the techniques of Theorems 3.1 and 7.1.

THEOREM 2. *Let f_n be estimate* (2), *where K is a bounded density with integrable radial majorant. The conditions*

$$\lim_{n \to \infty} h_n = 0; \qquad \lim_{n \to \infty} nh_n^d = \infty \tag{10}$$

are sufficient for the weak convergence to 0 *of* $|f_n - f|$, *almost all* x. *If also*

$$\lim_{n \to \infty} \frac{nh_n^d}{\log \log n} = \infty, \tag{11}$$

then $|f_n - f| \to 0$ *almost surely, almost all* x. *Finally, if in addition*

$$\lim_{n \to \infty} \frac{nh_n^d}{\log n} = \infty,$$

then $|f_n - f| \to 0$ *completely, almost all* x.

Proof. Let us introduce the random variables $Y_n = h_n^{-d}K((x - X_n)/h_n)$. By Lemma 2, $E(Y_n) = f * K_{h_n} \to f$, almost all x, if $h_n \to 0$. Now, $E(f_n) = (1/n)\sum_{i=1}^{n} E(Y_i) \to f$, almost all x, by Toeplitz's lemma (Hall and Heyde,

1980, p. 31). Let $m_n = \inf_{i \le n} h_i^d$, and let M be the supremum of $K(x)$ over all x. Let c be $\sup_i E(Y_i)$ (this depends upon x). Note that $E(Y_i^2) \le cM/h_i^d$ and $|Y_i - E(Y_i)| \le M/h_i^d$. For arbitrary $\varepsilon > 0$, we have by Bennett's inequality, also used in the proof of Lemma 6.5, and for all x for which $c < \infty$,

$$P(|f_n(x) - E(f_n(x))| \ge \varepsilon) = P\left(\left|\frac{1}{n}\sum_{i=1}^{n}(Y_i - E(Y_i))\right| \ge \varepsilon\right)$$

$$\le 2\exp\left(-\frac{n\varepsilon^2}{2(cM/m_n + \varepsilon M/m_n)}\right),$$

and this tends to 0 for all $\varepsilon > 0$ because $nm_n \to \infty$, which is implied by (10) (see Lemma 3 given immediately following this proof). This weak convergence is thus valid for almost all x, and the first part of the theorem follows.

For the strong convergence, we will apply a version of the strong law of large numbers (Loève, 1963, p. 253), which asserts that if $|Y_n| \le an$ for all n and some $a < \infty$ (which is the case here since the h_n are positive numbers and $nh_n^d/(\log\log n) \to \infty$), then $(1/n)\sum_{i=1}^{n}(Y_i - E(Y_i)) \to 0$ almost surely if and only if for all $\varepsilon > 0$,

$$\sum_{k=0}^{\infty} P\left(\left|2^{-k}\sum_{i=2^k+1}^{2^{k+1}}(Y_i - E(Y_i))\right| \ge \varepsilon\right) < \infty$$

(this is also called Prohorov's convergence criterion (1949)). But, again by Bennett's inequality, we can conclude that $f_n - E(f_n) \to 0$ almost surely for almost all x when

$$\sum_{k=0}^{\infty} 2\exp\left(-\frac{2^k\varepsilon^2}{2(cM/m_{2^k} + \varepsilon M/m_{2^k})}\right) < \infty,$$

which is implies by $2^k m_{2^k}/\log k \to \infty$ as $k \to \infty$. This in turn follows from $nm_n/\log\log n \to \infty$ and thus from $nh_n^d/\log\log n \to \infty$ (Lemma 3). This concludes the proof of Theorem 2.

LEMMA 3. *If $a_n, b_n \ge 0$, $a_n \uparrow \infty$, then $a_n/b_n \to \infty$ if and only if $a_n/\sup_{i \le n} b_i \to \infty$.*

Proof. Note that

$$\frac{a_n}{b_n} \ge \frac{a_n}{\sup_{i \le n} b_i} \ge \min\left(\inf_{i > N}\frac{a_i}{b_i}, \frac{a_n}{\sup_{i \le N} b_i}\right).$$

Lemma 3 follows by first picking N large enough and then letting n grow unbounded.

4. MAXIMUM LIKELIHOOD ESTIMATES

The classical maximum likelihood estimation principle, when applied here, would be such that the estimate f_n is the density g that maximizes

$$\prod_{i=1}^{n} g(X_i). \tag{12}$$

This maximum is not achievable without restrictions on the class of allowable g. Roughly speaking, we would approach a discrete distribution with atoms at the X_i's. There are several cures for this problem. For example, Grenander (1981) suggests a remedy for this by choosing g from an appropriate collection of densities C_n, which is allowed to grow slowly with n. The sequence of collections C_n is called a *sieve* and the resulting estimation method is called the *method of sieves*. Several examples of sieves are given below. In the last example, we relate the method of sieves to other well-known maximum likelihood density estimates.

EXAMPLE 1 (The Histogram Estimate). If

$$C_n = \{ g\colon g \text{ is constant on } [(j-1)h_n, jh_n),\ j \text{ integer} \},$$

where $h_n \to 0$ as $n \to \infty$, then (12) is maximized by the fixed histogram estimate

$$f_n(x) = \frac{1}{h_n} \frac{1}{n} \sum_{i=1}^{n} I_{[X_i \in [(j-1)h_n, jh_n)]}, \quad x \in [(j-1)h_n, jh_n)$$

(see Section 3.2 of Tapia and Thompson, 1978).

EXAMPLE 2 (The Convolution Sieve). Geman and Hwang (1982) suggest the sieve

$$C_n = \{ g\colon g = K_{h_n} * \nu \quad \text{for some probability measure } \nu \},$$

where $h_n \to 0$ as $n \to \infty$, and K is the normal density. C_n is called a *convolution sieve*. There is no particular reason to take the normal density except perhaps for theoretical convenience. For example, maximizing (12) for the normal convolution sieve gives an estimate of the form

$$f_n(x) = \sum_{i=1}^{n} p_i \frac{1}{h_n} K\left(\frac{x - y_i}{h_n}\right),$$

for some probability vector $(p_1, \ldots, p_n)$ and some real numbers $y_1, \ldots, y_n$ all strictly contained in $(\min X_i, \max X_i)$ (this was shown by Geman and McClure, and the proof can be found in Geman (1981)). We note that although the Parzen–Rosenblatt kernel estimate is in the sieve, it is not among the optimal solutions.

The computation of the optimal values $y_1, \ldots, y_n, p_1, \ldots, p_n$ is difficult. If $p_i = 1/n$ for all i, a solution is somewhat easier to obtain, but it should be clear that computational difficulties are inherent in all the maximum likelihood based estimates. The fact that explicit solutions are not available makes the analysis difficult too. Rate of convergence results are nonexistent, but some consistency results can be found in the literature. For example, for the normal convolution sieve

$$C_n = \left\{ g\colon g(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - y_i}{h_n}\right), \text{ some } y_1, \ldots, y_n \in R^1 \right\},$$

Geman (1981) showed that

$$\sup_{\substack{\text{all optimal} \\ \text{solutions of (12)} \\ \text{in } C_n}} \int |f_n - f| \to 0 \quad \text{almost surely} \tag{13}$$

under the following conditions: $h_n \to 0$, $n^a h_n \to \infty$ for some $0 < a < \frac{1}{4}$, f has compact support, and

$$\int f \log f < \infty. \tag{14}$$

One of the conditions here states that we can't let C_n grow too large too quickly. The condition on the peakedness of f (see (14)) comes naturally because maximizing (12) is equivalent to maximizing $\sum_{i=1}^{n} \log g(X_i)$, which has expected value $n \int f \log g$. Now, in view of

$$\int f \log g \le \int f \log f, \quad \text{all } f, g$$

(which is a consequence of Jensen's inequality), there is hope that maximizing (12) gives us a g that is close to f, at least if some law of large numbers applies, that is, when (14) holds. In this sense, (14) is not a natural criterion. The only positive thing about it is that there is some vague one-way

connection with the L_1 error:

$$\sqrt{2\left(\int f \log f - \int f \log g\right)} \geq \int |f - g|$$

(see Theorem 8.2).

Suppose now that we also allow h_n to act as a parameter in the convolution sieve. Then, the maximum of (12) would be achieved by setting $h_n = 0$, $y_i = X_i$, all i. Thus, if we are going to maximize over h_n too, another device is needed, for example, a limitation of the complexity of the mixture as suggested by Geman and Hwang (1982), who define the sieve

$$C_n = \left\{ g\colon g(x) = \frac{1}{k_n} \sum_{i=1}^{k_n} \frac{1}{h} K\left(\frac{x - y_i}{h}\right) \text{ for some } h > 0,\ y_1, \ldots, y_{k_n} \in R^1 \right\}.$$

For this sieve, statement (13) remains valid when f is bounded and has compact support, and when $k_n \to \infty$ and $k_n/n^a \to 0$ as $n \to \infty$ for some $a < 1/5$ (Geman and Hwang, 1982). Unfortunately, we are still left with the problem of choosing k_n.

EXAMPLE 3 (The Method of Penalized Maximum Likelihood). Let $\mathcal{G}$ be a suitable class of densities, and let C_n now be defined by

$$C_n = \{ g\colon g \in \mathcal{G}, \Psi(g) \leq M \},$$

where m is a constant possibly depending upon n, and $\Psi(g)$ is a penalty function penalizing for oscillatory behavior. This sieve method is suggested but not analyzed in Geman and Hwang (1982). The Lagrange multiplier method corresponding to it would find the g in $\mathcal{G}$ that maximizes

$$\sum_{i=1}^{n} \log g(X_i) - \lambda_n \Psi(g), \tag{15}$$

where $\lambda_n \geq 0$ is a Lagrange multiplier. This multiplier plays the role of a smoothing factor. There are several suggestions for Ψ and $\mathcal{G}$, for example,

(i) $\Psi(g) = \int g'^2/g$; $\mathcal{G} = \{ g\colon (\sqrt{g})' \in L_2 \}$ (Good and Gaskins, 1971; see also Tapia and Thompson, 1978, pp. 108–109);

(ii) $\Psi(g) = a\int g'^2 + b\int g''^2$; $\mathcal{G} = \{ g\colon (\sqrt{g})' \in L_2, (\sqrt{g})'' \in L_2 \}$; $a \geq 0$, $b > 0$ (Good and Gaskins, 1971).

The main problem, once again, is computational: how does one find these

maxima? Some solution based upon a quantization of the sample space is given in Scott et al. (1980). Unfortunately, even though their estimate is consistent for large classes of densities f, the experimentally obtained asymptotic error rate in L_2 is slightly worse than that of the standard kernel estimate in some simple problems.

For consistency, we must let λ_n vary with n such that $\lambda_n \to 0$. The rate must be controlled, because for $\lambda_n = 0$, we obtain a degenerate solution. A variety of conditions on Ψ and f guaranteeing consistency are given in de Montricher (1980), Klonias (1982), and Silverman (1982).

5. VARIABLE HISTOGRAM ESTIMATES

The histogram estimate studied in Chapters 3 and 5 is not locally sensitive: the size of the cells is not allowed to vary with x. They are but special cases of *variable histogram estimates* defined as follows:

(i) Determine a countable (possibly finite) partition $P_{1n}, P_{2n}, \ldots$ of R^d. This partition is allowed to depend upon the data $X_1, \ldots, X_n$.

(ii) Estimate f on P_{in} by a constant c_{in} such that the estimate itself is a density, that is, $c_{in} \geq 0$, all i, and $\sum_i c_{in}\lambda(P_{in}) = 1$ (λ is Lebesgue measure). Usually, but not necessarily, $c_{in} = N_{in}/n\lambda(P_{in})$, where N_{in} is the number of data points falling in P_{in}.

Thus, the kernel estimate with a uniform $[-1, 1]$ kernel K classifies as a variable histogram estimate. But more importantly, the class of variable histogram estimates is large enough to allow the asymptotic L_1 error rate $n^{-2/5}$ for some f. The original fixed grid histogram estimate had a built-in limitation of $n^{-1/3}$.

Among the variable histogram estimates, perhaps the most popular type of estimate is that based upon statistically equivalent blocks. In R^1, for example, we could consider the order statistics $X_{(1)}, \ldots, X_{(n)}$ corresponding to the data, and partition the space by defining $P_{1n} = [X_{(1)}, X_{(k)}]$, $P_{2n} = (X_{(k)}, X_{(2k)}], \ldots$, so that each interval has about k data points, or is empty. For all points x in $[X_{(1)}, X_{(n)}]$, estimate $f_n(x)$ by $k/n\lambda(P_{in})$, $x \in P_{in}$.

This estimate is (not explicitly) suggested in Anderson (1965), and formally defined and studied by Van Ryzin (1970, 1973). Smooth spline functions that generalize it (but possibly violate the density property) can be found in Wahba (1971, 1975, 1976). Its L_1 consistency was first obtained under very general conditions by Abou-Jaoude (1976), and its L_1 rate of convergence was studied by Hanna and Abou-Jaoude (1981).

THEOREM 3 (Abou-Jaoude, 1976). *For the order statistics based histogram estimate with k data points per interval, the following statements are*

equivalent:

A. $\int |f_n - f| \to 0$ *is probability for all Riemann integrable* f;

B. $\int |f_n - f| \to 0$ *completely for all Riemann integrable* f;

C. $\lim_{n\to\infty} k = \infty$ *and* $\lim_{n\to\infty}(k/n) = 0$.

There are many possible generalizations to R^d, one of which is given in Gessaman (1970), where the first axis is cut into about $(n/k)^{1/d}$ intervals each containing about an equal number of data points (this will be called a homogeneous cut). Each of the cylindrical sets defined by these intervals is subjected to another homogeneous cut into $(n/k)^{1/d}$ pieces, but now along the second axis. After d homogeneous cuts, each of the n/k final "cells" has about k data points. Gessaman (1970) offers some pointwise consistency results, and points out quite correctly that the computation of f_n after the construction of the partition is fast.

Alternatively, one could cut each axis in turn, on a rotational basis, each time splitting the remaining data points exactly in half, and stopping when all the cells have about k data points. This method has the computational advantage that the implementation could be done with a balanced binary tree of about $\log_2(n/k)$ levels.

6. KERNEL ESTIMATES WITH REDUCED BIAS

Variations of the standard density estimates can, in some cases, give better rates of convergence for $E(J_n)$ than those obtained in Chapter 5 for the kernel and histogram estimates. Usually, the improvement is due to a reduction in the bias component $\int |E(f_n) - f|$ and is possible only for very smooth f. In this section. We will illustrate some general bias reduction principles and illustrate them for the kernel estimate.

In what follows we will assume that f_n is a density estimate of f (f_n itself is a density in x), and that g_n is another function on R^d. An *additive variation* of f_n can be defined as follows:

$$f_n^* = \frac{(f_n + g_n)_+}{\int (f_n + g_n)_+}; \quad \int g_n = 0; \quad \int |g_n| < \infty, \quad \text{all } n.$$

The normalization is necessary to insure that f_n^* is a density. A *multiplica-*

tive variation of f_n is defined by

$$f_n^* = \frac{f_n g_n}{\int f_n g_n}; \quad g_n \geq 0, \quad \text{all } n.$$

For L_1 convergence, the normalizations can be ignored because of the following lemma:

LEMMA 4. *For all densities f, and all density estimates f_n on R^d, any additive variation satisfies*

$$\int |f_n^* - f| \leq \int |(f_n + g_n) - f|.$$

Similarly, any multiplicative variation of f_n satisfies

$$\int |f_n^* - f| \leq \int |f_n g_n - f| + \left| \int f_n g_n - 1 \right|.$$

Proof. For the additive variation, we refer to the nonnegative projection Theorem 11.4. For the multiplicative variation, we argue as follows: when $\int f_n g_n \leq 1$, we have $f_n^* \geq f_n g_n$, and thus,

$$\int |f_n^* - f| = 2 \int (f - f_n^*)_+ \leq 2 \int (f - f_n g_n)_+$$

$$= \int |f - f_n g_n| + \left(1 - \int f_n g_n\right).$$

When $\int f_n g_n \geq 1$, then $f_n^* \leq f_n g_n$ and, therefore,

$$\int |f_n^* - f| = 2 \int (f_n^* - f)_+ \leq 2 \int (f_n g_n - f)_+$$

$$= \int |f_n g_n - f| + \left(\int f_n g_n - 1\right).$$

The foremost example of bias reduction is Bartlett's estimate (1963) on R,

$$f_n(x) = (nh)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where K is a Borel measurable function with the following properties:

(i) K is symmetric, bounded, and has compact support;

(ii) $\int K = 1$;

(iii) $\int x^i K = 0,\ i = 1, 2, \ldots, s-1$; **(16)**

(iv) $\int |x|^s |K| < \infty$;

where $s \geq 1$ is a fixed integer. For $s > 1$, f_n can possibly take negative values since K does. It is easy to verify that f_n can be written as a standard kernel estimate (with kernel $K_+/\int K_+$) plus a function g_n with zero integral. Thus the density $f_n^* = (f_n)_+/\int (f_n)_+$ qualifies as a genuine additive variation of the standard kernel estimate. To avoid ambiguity, we will call f_n *Bartlett's estimate* and f_n^* the *normalized Bartlett estimate*. For a different point of view on this estimate, see Section 5.9.

THEOREM 4. *Assume that $h \to 0$ and $nh \to \infty$. Then, for any integer $s \geq 1$, Bartlett's estimate is exponentially convergent, that is, for all $\varepsilon > 0$, there exist positive numbers r and n_0 such that*

$$P\left(\int |f_n - f| > \varepsilon\right) \leq e^{-rn}, \quad n \geq n_0.$$

This property holds true for all f, but r can be chosen independently of f. Theorem 4 remains valid for the normalized Bartlett estimate.

Proof. Theorem 4 is a direct consequence of Theorem 3.1, where K was subject to only two conditions: $\int K = 1$, $\int |K| < \infty$. For the normalized Bartlett estimate, use Lemma 4.

THEOREM 5. *Assume that $s \geq 1$ is a fixed integer and that $h \to 0$, $nh \to \infty$. Let f be a density with compact support and (2s-1) absolutely continuous derivatives, and let $f^{(2s)}$ be continuous. In the notation of Chapter 5, we have, both for Bartlett's estimate and the normalized Bartlett estimate,*

$$E(J_n) \sim \int \frac{\alpha\sqrt{f}}{\sqrt{nh}} \psi\left(\frac{h^{2s}}{(2s)!}\beta_{2s}|f^{(2s)}|\right)$$

$$\leq \frac{\alpha \int \sqrt{f}}{\sqrt{nh}} \sqrt{\frac{2}{\pi}} + \frac{h^{2s}}{(2s)!}\beta_{2s} \int |f^{(2s)}|.$$

Here $\alpha = \sqrt{\int K^2}$ *and* $\beta_{2s} = \int |x|^{2s} K \neq 0$. *Furthermore,*

$$\limsup_{n \to \infty} \inf_h n^{2s/(4s+1)} E(J_n) \leq C_{2s} A_{2s}(K) D_{2s}(f),$$

where

$$C_{2s} = \frac{1 + 4s}{(2s)!} \left(\frac{2}{\pi} \right)^{2s/(4s+1)} \left(\frac{(2s)!}{4s} \right)^{4s/(4s+1)},$$

$$A_{2s}(K) = \left(\alpha^{4s} |\beta_{2s}| \right)^{1/(4s+1)} = \left(\left(\int K^2 \right)^{2s} \left| \int x^{2s} K \right| \right)^{1/(4s+1)},$$

and

$$D_{2s}(f) = \left(\left(\int \sqrt{f} \right)^{4s} \int |f^{(2s)}| \right)^{1/(4s+1)}.$$

This upper bound is not exceeded for the choice

$$h = \left(\sqrt{\frac{2}{\pi}} \frac{\alpha (2s-1)!}{2|\beta_{2s}|} \frac{\int \sqrt{f}}{\int |f^{(2s)}|} \right)^{2/(4s+1)} n^{-1/(4s+1)}.$$

If $\beta_{2s} = 0$, *and* $D_{2s}(f) < \infty$, *then*

$$\limsup_{n \to \infty} \inf_h n^{2s/(4s+1)} E(J_n) = 0.$$

If $\beta_2 = 0$ *and* f *is any density with compact support and* $B^*(f) < \infty$ (*notation of Theorem* 5.1), *then*

$$\limsup_{n \to \infty} \inf_h n^{2/5} E(J_n) = 0.$$

The quantity $D_{2s}(f)$ appearing in the upper bound of Theorem 5 can also be found in the minimax lower bounds of Theorems 4.2 and 4.3. More importantly, for individual f with finite values of $D_{2s}(f)$ we can do much better than the rate $n^{-2s/(4s+1)}$ given by the minimax lower bound, simply by insuring that the kernel K satisfies (16) *and* has $\beta_{2s} = \int x^{2s} K = 0$. This apparent contradiction can be explained by the fact that the improvement is not uniform over the class of all f with compact support and $D_{2s}(f) \leq r$ for fixed constant r. In fact, within this class, all slow rates to zero are

achievable for $n^{2s/(4s+1)}E(J_n)$. The last half of Theorem 5 does not provide us with any clues as to how h should be chosen either. From the proof, it is clear that $h = Mn^{-1/(4s+1)}$ is better than $h = M^*n^{-1/(4s+1)}$ whenever $M > M^*$, but that is about all one can conclude without further assumptions about f. A similar problem must be faced for K, and it seems sensible to choose h and K in such a way that a minimax bound (such as the one obtained in Theorem 5.12) is minimized. For example, we could also obtain in Theorem 5 a crude but manageable upper bound merely by replacing β_{2s} throughout by $\int x^{2s}|K|$ (i.e., in the definition of h and of $A_{2s}(K)$), and choose K so that $A_{2s}(K)$ is minimal. Let us mention here that the following kernels satisfy the conditions (16)(ii), (iii), and $\int x^{2s}K = 0$ for $s = 1$:

$$K(x) = \tfrac{9}{8}\left(1 - \tfrac{5}{3}x^2\right), \qquad |x| \le 1 \quad \text{(Bartlett, 1963)},$$

$$K(x) = \tfrac{3}{2}\left(1 - \tfrac{1}{3}x^2\right)e^{-x^2/2} \quad \text{(Rosenblatt, 1971)}.$$

Other such kernels are of course easy to construct too (Deheuvels, 1977).

The choice of a kernel is determined by the degree of smoothness that we expect to observe in f. There are admittedly other considerations too. For example, some applications demand the simultaneous estimation of f and one or more of its derivatives. If the derivatives of f are estimated by the corresponding derivatives of the estimate f_n, then it is obvious that K should be smooth. This point was addressed by Müller (1984) and Gasser et al. (1983). Finally, we note that if h is taken in accordance with Theorem 5 and f does not have the smoothness properties called for by the theorem, one could actually lose quite a bit in asymptotic performance.

Proof of Theorem 5. The essential difference with Theorem 5.1 is in the bias term. By Taylor's expansion of f about x,

$$f(y) = \sum_{i=0}^{2s} (y-x)^i \frac{f^{(i)}(x)}{i!} + \frac{(y-x)^{2s}}{(2s)!}\left(f^{(i)}(\xi) - f^{(i)}(x)\right),$$

$$y \ge \xi \ge x,$$

and a symmetric expression when $y < x$, we obtain

$$\left| \int \frac{1}{h} K\left(\frac{x-y}{h}\right)\left(f(y) - f(x)\right) dy - \int \frac{h^{2s}}{(2s)!}\beta_{2s} f^{(2s)}(x) \right|$$

$$\le \int \frac{1}{h} K\left(\frac{x-y}{h}\right) \frac{|y-x|^{2s}}{(2s)!} |f^{(2s)}(\xi) - f^{(2s)}(x)|\, dy,$$

where ξ depends upon x, y, and f. In other words, because of property (16), all the terms in the Taylor series expansion cancel out after convolution with K_h except the $2s$-th term. The bound on the right-hand side is $o(h^{2s})$ uniformly over all x in a large interval T (by the uniform continuity of $f^{(2s)}$ and the compactness of the support of K), and 0 outside T. If we define, as in Chapter 5, $B_n = E(f_n) - f$, then this implies

$$\int |B_n| - \frac{h^{2s}}{(2s)!}|\beta_{2s} f^{(2s)}|| = o(h^{2s}),$$

a result that will replace Lemma 5.11.

Lemma 5.10 remains obviously valid, and thus Theorem 5.1 can be followed to the letter if we replace z there by $h^{2s}|\beta_{2s}|f^{(2s)}|/(2s)!$. This proves the first half of Theorem 5.

We note that the asymptotic upper bound is of the form $uh^{-1/2} + vh^{2s}$ for some positive numbers u, v not depending upon h. Thus, a formal minimization with respect to h gives the following minimal value:

$$(u^{4s}v)^{1/(4s+1)} \frac{(4s+1)}{(4s)^{4s/(4s+1)}}.$$

It is attained for $h = (u/4sv)^{2/(4s+1)}$. If we replace u by $\alpha \int \sqrt{f} \sqrt{2/\pi n}$ and v by $|\beta_{2s}| \int |f^{(2s)}|/(2s)!$, we obtain the announced upper bound.

For the case $\beta_{2s} = 0$, $D_{2s}(f) < \infty$, take M arbitrarily large, set $h = Mn^{-1/(4s+1)}$, and note that $\int |B_n| = o(n^{-2s/(4s+1)})$. Also, in the notation of Lemma 5.10,

$$\frac{\int \sigma_n}{\sqrt{\int K^2}} \leq \frac{\int \sqrt{f} + o(1)}{\sqrt{nh}} \leq \frac{\int \sqrt{f} + o(1)}{\sqrt{M}} n^{-2s/(4s+1)}.$$

This concludes the proof of the theorem.

As a second example, partially overlapping with the previous example, we mention the time-honored *jackknife method* for reducing the bias of estimates in statistics (Quenouille, 1956; see also Schucany et al., 1971). For density estimation, it was first developed and illustrated in Sommers (1972) and Schucany and Sommers (1977).

Let $f_{n1}, \ldots, f_{nM}$ be M kernel estimates of f, all based upon the same sample $X_1, \ldots, X_n$, but possibly with different kernels $K_1, \ldots, K_M$ and

different smoothing factors $h_1, \ldots, h_M$. We will assume that $h_i = ha_i$ for some constants a_i; h depends upon n only. Consider now the linear combination

$$f_n = \frac{\sum_{i=1}^{M} b_i f_{ni}}{\sum_{i=1}^{M} b_i},$$

where $b_1, \ldots, b_M$ are constants not summing to 0. Clearly, if all the kernels integrate to 1, so does f_n. Thus, its normalized form is again an additive variation of the standard kernel estimate. It is not hard to verify that f_n coincides with Bartlett's estimate with kernel

$$K(x) = \frac{\sum_{i=1}^{M} b_i((1/a_i)K_i(x/a_i))}{\sum_{i=1}^{M} b_i},$$

and Theorems 4 and 5 apply. Now, if all the K_i's satisfy Bartlett's condition (16) (i) *only* (and not (ii), (iii), and (iv)), and if all the K_i's are densities, and if f satisfies the conditions of Theorem 5, we have the following Taylor series expansion for the pointwise bias:

$$E(f_{ni}) - f = \sum_{j=1}^{s} \frac{h^{2j}}{(2j)!} f^{(2j)} \int x^{2j} K_i + o(h^{2s}).$$

The first $s - 1$ terms in the bias of f_n can be eliminated if

$$\sum_{i=1}^{M} b_i a_i^{2j} \int x^{2j} K_i = 0, \qquad j = 1, 2, \ldots, s-1.$$

This system of equations has many degrees of freedom. For example, we could take all K_i's equal to K, and set the a_i's equal to i. Then the equations can be reduced to

$$\sum_{i=1}^{M} b_i i^{2j} = 0, \qquad j = 1, 2, \ldots, s-1.$$

This has nonzero solutions for the b_i's whenever $M \geq s$. For example, with $s = M = 2$, we have a solution $b_1 = 1$, $b_2 = -\frac{1}{4}$, and a *jackknife estimate*

$$f_n = \tfrac{1}{3}(4f_{n1} - f_{n2}) = f_{n1} + \tfrac{1}{3}(f_{n1} - f_{n2}),$$

which seems to suggest that $\frac{1}{3}(f_{n1} - f_{n2})$ is a correction factor of sorts for the standard kernel estimate f_{n1}.

As an example of a multiplicative variation on the standard kernel estimate, and certainly not the only possible one, we will present the *estimate of Terrell and Scott* (1980). Consider two kernel estimates with the same symmetric bounded compact support kernel K, and smoothing factors h and $2h$, respectively: f_{n1}, f_{n2}. Again, both estimates are based upon the same sample $X_1, \ldots, X_n$. Then form the estimate

$$f_{n1}\left(\frac{f_{n1}}{f_{n2}}\right)^{1/3}.$$

The multiplicative correction factor is always well defined if we ensure that K is unimodal, in which case it is a number between 0 and $2^{1/3}$. The consistency of this estimate requires some work, for we must establish that

$$\int \left| f_{n1}\left(\frac{f_{n1}}{f_{n2}}\right)^{1/3} - f \right| \to 0.$$

But the left-hand side is bounded from above by

$$2^{1/3}\int |f_{n1} - f| + \int f \left| \left(\frac{f_{n1}}{f_{n2}}\right)^{1/3} - 1 \right|.$$

Thus, by Theorems 3.1 and 6.3 and the Lebesgue dominated convergence theorem, we can conclude the following:

THEOREM 6. *Let K be a symmetric bounded unimodal density on R with compact support, and let $h \to 0$, $nh \to \infty$. Then, for* all *densities f,*

$$\int |f_n^* - f| \to 0 \quad \textit{in probability as } n \to \infty,$$

where f_n^* is the normalized form of the Terrell and Scott estimate. The same is true for all correction factors g_n that are uniformly bounded from below by 0 *and from above by a positive constant, and that converge pointwise almost everywhere to* 1.

There is also an almost sure version of Theorem 6, requiring only that the condition $nh/(\log\log n) \to \infty$ be added. Terrell and Scott have shown that

under sufficient smoothness conditions for f, the bias is $O(h^4)$ at each point x, while the variance is $O((nh)^{-1})$ at each point x. The estimate seems to behave asymptotically as Bartlett's estimate with $s = 2$, but a rigorous analysis is lacking at this point. The main advantage of the multiplicative variations over the additive variations is their nice behavior in the tails: additive variations of the kernel estimate will often abruptly drop to 0 because of the normalization. The normalization of a multiplicative variation of the kernel estimate is less drastic.

7. GRENANDER'S ESTIMATE FOR MONOTONE DENSITIES

In this section, we will consider only monotone densities on $[0, \infty)$ (the class of all these densities will be called M), and monotone densities on $[0, 1]$ with $f(0) \leq B$ (the collection of these densities will be called M_B). We have seen that the minimax lower bound over M is at least $\frac{1}{8}$ (Theorem 4.1), but that the minimax lower bound over M_2 is $(\frac{1}{16} + o(1))(4/n)^{1/3}$ (Theorem 4.9). Lucien Birgé has proved that there exists a minimax lower bound over M_B of $0.198(\log(B + 1)/n)^{1/3}$, valid for $B \geq 1.3$, $B/n \leq 0.026$ (since this is an unpublished result, it is not included in this book). With these results in mind, we can now compare various estimators. There is no reason of course to look at estimates that are consistent for all f if one is only interested in M or M_B. One such estimate is *Grenander's maximum likelihood estimate* (Grenander, 1956): it is a density in M for which the product

$$\prod_{i=1}^{n} f_n(X_i)$$

is maximal. This optimization problem has a remarkably simple solution: if $F_n(x) = (1/n)\sum_{i=1}^{n} I_{[X_i \leq x]}$ is the empirical distribution function, and $G_n(x)$ is the smallest concave majorant of F_n (i.e., G_n is obtained by taking a huge elastic band, putting it around the first quadrant, and letting it go: it will come to rest, if we hold it at the x axis, around the curve of F_n, and is of course piecewise linear), then $f_n = G_n'$.

For this estimate, a deep analysis of its pointwise properties can be found in Prakasa Rao (1969). But the result that interests us most of all, and is shockingly beautiful, is due to Groeneboom (1983):

$$n^{1/6}\left(n^{1/3}\int |f_n - f| - C(f)\right) \xrightarrow{\mathcal{L}} N(0, \sigma^2),$$

where $\xrightarrow{\mathscr{D}}$ means "converges in distribution to", $N(0, \sigma^2)$ is a normal $(0, \sigma^2)$ random variable, $\sigma > 0$ is a constant independent of f, and

$$C(f) = c_0 \int_0^1 \left(\frac{1}{2}|f'|f\right)^{1/3},$$

where c_0 is another universal constant (its value is about 0.82, based upon experimental results reported by Groeneboom). Groeneboom's result is valid for all strictly decreasing f on $[0,1]$ with continuous and bounded second derivative, and $f' < 0$ on $(0,1)$. For these f, we have

$$n^{1/3} E\left(\int |f_n - f|\right) \to c_0 \int \left(\frac{1}{2}|f'|f\right)^{1/3}.$$

Grenander's estimate performs even better when f has flat parts. For the uniform density on $[0,1]$, rate $n^{-1/2}$ is achieved.

Let us compare this result for individual f with results for the kernel estimate. For the densities f covered by Groeneboom's theorem,

$$\int \left(\frac{1}{2}|f'|f\right)^{1/3} \le \left(\frac{1}{2}\left(\int \sqrt{f}\right)^2 \left(\int |f'|\right)\right)^{1/3} = B_H(f).$$

For the kernel estimate with h chosen as in Theorem 5.10 we know that $n^{1/3} E(\int |f_n - f|) \le (1.24 \cdots + o(1)) B_H(f)$. Thus, this seems slightly worse than Grenander's estimate. On the other hand, for the kernel estimate, suitably modified near zero, the rate $n^{-1/3}$ is guaranteed for all f in M_B:

THEOREM 7. *Let f be any density in M_B, and let h be chosen as follows:*

$$h = \left(\frac{6}{\pi n B^2}\right)^{1/3}.$$

Let K be the isosceles triangular density on $[-1,1]$. Let g_n be the kernel estimate obtained from $Y_1, \ldots, Y_n$, where Y_i is equal to X_i with a random sign added (by a coin flip), and let f_n be defined by

$$f_n(x) = g_n(x) + g_n(-x), \qquad x > 0.$$

Then

$$\limsup_{n \to \infty} n^{1/3} E\left(\int |f_n - f|\right) \le \left(\frac{6}{\pi}\right)^{1/3} B^{1/3}.$$

Proof. We have

$$\int |f_n - f| \le \int_0^\infty \left| g_n(x) - \frac{1}{2} f(x) \right| dx + \int_0^\infty \left| g_n(-x) - \frac{1}{2} f(x) \right| dx$$

$$\le \int_{-\infty}^\infty \left| g_n(x) - \frac{1}{2} f(|x|) \right| dx.$$

We apply Theorem 5.10 directly to the latter L_1 error. The reason for the rather artificial symmetrization is the following: we want to ensure that $B_H^*(f)$ is uniformly bounded over our class of densities. Now, if take ϕ in the definition of $B_H^*(f)$ symmetric and unimodal, then it is not at all sure that $f * \phi_a$ is unimodal. However, if $g(x) = \frac{1}{2} f(|x|)$, then $g * \phi_a$ is indeed unimodal (Feller, 1971), so that $\int |(g * \phi_a)'| = 2g(0) = B$, for all a. We have, in view of $\int \sqrt{g} = \sqrt{2} \int \sqrt{f}$, $B_H^*(g) = ((\int \sqrt{f})^2 B)^{1/3} \le B^{1/3}$. The remainder of the proof follows directly from Theorem 5.10.

The situation is even rosier for the kernel estimate, because, for the densities covered by Groeneboom's theorem, we can choose h and K in such a way that $n^{2/5} E(\int |f_n - f|)$ tends to a constant. For this, it is absolutely necessary to use the symmetrization trick of Theorem 7 (for otherwise, this result would be impossible by the discontinuity at zero). In other words, on an individual basis, a suitably modified kernel estimate can be much better than Grenander's estimate.

We also have the following minimax upper bound, simply because it is achievable by the modified kernel estimate:

THEOREM 8. *For the estimate f_n of Theorem 7, and all $B \ge 1$,*

$$\limsup_{n \to \infty} n^{1/3} \sup_{f \in M_B} E\left(\int |f_n - f| \right) \le \left(\frac{6}{\pi} \right)^{1/3} B^{1/3}.$$

Proof. It suffices to verify that the $o((nh)^{-1/2})$ term in Theorem 5.10 remains $o((nh)^{-1/2})$ uniformly over the class of symmetric unimodal densities on $[-1, 1]$ that are bounded by $B/2$. For the remainder of the proof, we refer to the proof of Theorem 7.

Unfortunately, this upper bound is not the best possible, even though for M_2, the ratio between minimax upper and lower bounds is only about 16. Birgé, in a private communication, has told us that the minimax upper bound for M_B is smaller than $1.98(\log(B + 1)/n)^{1/3}$, valid for $B \ge 1.3$, $B/n \le 0.026$. It seems very likely that this better minimax upper bound is

attainable with Grenander's estimate. This belief is based upon the following observation:

LEMMA 5. *Let f be an absolutely continuous density in M_B with almost everywhere derivative f'. Then*

$$\int (|f'|/f)^{1/3} \leq 1 + (\log B)^{1/3}.$$

Proof. Let $u \in [0, 1]$ be a point such that $f(x) \geq 1$, $x < u$, and $f(x) \leq 1$, $x > u$. Then, by Jensen's inequality,

$$\begin{aligned}
\int (|f'|/f)^{1/3} &\leq \int_0^u (|f'|/f)^{1/3} + \int_u^1 |f'|^{1/3} \\
&\leq \int_0^u f\left(\frac{|f'|}{f^2}\right)^{1/3} + f(u)^{1/3} \\
&\leq \left(\int_0^u f\left(\frac{|f'|}{f^2}\right)\right)^{1/3} + 1 \\
&= \left(\int_0^u -d(\log f)\right)^{1/3} + 1 \\
&= (\log B)^{1/3} + 1.
\end{aligned}$$

To close this section, we will merely give a few references to other estimates designed for the class of all unimodal densities: see, for example, Robertson (1967) and Wegman (1969, 1970a, 1975).

REFERENCES

S. Abou-Jaoude (1976). Sur la convergence L_1 et L_∞ de l' estimateur de la partition aléatoire pour une densité, *Annales de l'Institut Henri Poincaré B* **12**, pp. 299–317.

I. S. Abramson (1982). On bandwidth variation in kernel estimates—a square root law, *Annals of Statistics* **10**, pp. 1217–1223.

I. A. Ahmad and P. Lin (1976). Nonparametric sequential estimation of a multiple regression function, *Bulletin of Mathematical Statistics* **17**, pp. 63–75.

T. W. Anderson (1965). Some nonparametric multivariate procedures based on statistically equivalent blocks, in *Multivariate Analysis I*, P. R. Krishnaiah (Ed).

G. Banon (1976). Sur un estimateur non paramétrique de la densité de probabilité, *Revue de Statistique Appliquée* **24**, pp. 61–73.

M. S. Bartlett (1963). Statistical estimation of density functions, *Sankhya Series A* **25**, pp. 245–254.

L. Breiman, W. Meisel, and E. Purcell (1977). Variable kernel estimates of multivariate densities, *Technometrics* **19**, pp. 135–144.

R. J. Carroll (1976). On sequential density estimation, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **36**, pp. 136–151.

H. I. Davies (1973). Strong consistency of a sequential estimator of a probability density function, *Bulletin of Mathematical Statistics* **15**, pp. 49–53.

P. Deheuvels (1973a). Sur une famille d'estimateurs de la densité d'une variable aléatoire, *Comptes Rendus de l'Académie des Sciences de Paris* **276**, pp. 1013–1015.

P. Deheuvels (1973b). Sur l'estimation séquentielle de la densité, *Comptes Rendus de l'Académie des Sciences de Paris* **276**, pp. 1119–1121.

P. Deheuvels (1974). Conditions nécessaires et suffisantes de convergence ponctuelle presque sûre et uniforme presque sûre des estimateurs de la densité, *Comptes Rendus de l'Académie des Sciences de Paris* **278**, pp. 1217–1220.

P. Deheuvels (1977). Estimation nonparamétrique de la densité par histogrammes généralises, *Revue de Statistique Appliquée* **25**, pp. 5–42.

P. Deheuvels (1979). Estimation séquentielle de la densité, *Contrib. en Prob. y Est. Mat. Ens. de la Mat. y Análisis*, pp. 156–169, University of Granada.

G. M. de Montricher (1980). On the consistency of maximum penalized likelihood density estimation, Technical Report, Department of Mathematics, Rice University, Houston, Texas.

G. M. de Montricher, R. A. Tapia, and J. R. Thompson (1975). Nonparametric maximum likelihood estimation of probability densities by penalty function methods, *Annals of Statistics* **3**, pp. 1329–1348.

L. Devroye (1979). On the pointwise and the integral convergence of recursive kernel estimates of probability densities, *Utilitas Mathematica* **15**, pp. 113–128.

L. Devroye and C. S. Penrod (1982). The strong uniform convergence of multivariate variable kernel density estimates, Technical Report, Applied Research Laboratories, The University of Texas, Austin, Texas.

W. Feller (1971). *An Introduction To Probability Theory and Its Applications*, Wiley, New York.

T. Gasser, H.-G. Müller, and V. Mammitzsch (1983). Kernels for nonparametric curve estimation, Sonderforschungsbereich 123, Stochastische Mathematische Modelle, Preprint 210, Universität Heidelberg.

S. Geman (1981). Sieves for nonparametric estimation of densities and regressions, Reports in Pattern Analysis No. 99, Division of Applied Mathematics, Brown University, Providence, Rhode Island.

S. Geman and C.-R. Hwang (1982). Nonparametric maximum likelihood estimation by the method of sieves, *Annals of Statistics* **10**, pp. 401–414.

M. P. Gessaman (1970). A consistent nonparametric multivariate density estimator based on statistically equivalent blocks, *Annals of Mathematical Statistics* **41**, pp. 1344–1346.

I. J. Good and R. A. Gaskins (1971). Nonparametric roughness penalties for probability densities, *Biometrika* **58**, pp. 255–277.

U. Grenander (1956). On the theory of mortality measurement. Part II, *Skandinavisk Aktuarietidskrift* **39**, pp. 125–153.

U. Grenander (1981). *Abstract Inference*, Wiley, New York.

P. Groeneboom (1983). Estimating a monotone density, Proceedings of the Neyman-Kiefer Conference.

L. Györfi (1981). Strong consistent density estimate from ergodic sample, *Journal of Multivariate Analysis* **11**, pp. 81–84.

J. D. F. Habbema, J. Hermans, and J. Remme (1978). Variable kernel density estimation in discriminant analysis, in *COMPSTAT 78*, L. C. A. Corsten and J. Hermans (Eds.), Physica Verlag, Wien.

P. Hall and C. C. Heyde (1980). *Martingale Limit Theory and Its Application*, Academic Press, New York.

B. Hanna and S. Abou-Jaoude (1981). Sur la vitesse de convergence de l'estimateur de la partition aléatoire d'une densité de probabilité, *Publications de l'Institut de Statistique des Universités de Paris* **26**, pp. 51–67.

E. Isogai (1978). On strong consistency of a sequential estimator of probability density, *Science Reports of Niigata University A* **15**, pp. 25–33.

E. Isogai (1979). Strong consistency and optimality of a sequential density estimator, *Bulletin of Mathematical Statistics* **19**, pp. 55–69.

E. Isogai (1982). Strong uniform consistency of recursive kernel density estimators, *Science Reports of Niigata University A* **18**, pp. 15–27.

V. K. Klonias (1982). Consistency of two nonparametric maximum penalized likelihood estimators of the probability density function, *Annals of Statistics* **10**, pp. 811–824.

M. Loève (1963). *Probability Theory*, Van Nostrand, Princeton, New Jersey.

D. O. Loftsgaarden and C. P. Quesenberry (1965). A nonparametric estimate of a multivariate probability density function, *Annals of Mathematical Statistics* **28**, pp. 1049–1051.

Y. P. Mack and M. Rosenblatt (1979). Multivariate k nearest-neighbor density estimates, *Journal of Multivariate Analysis* **9**, pp. 1–15.

D. S. Moore and J. W. Yackel (1977). Consistency properties of nearest-neighbor density estimates, *Annals of Statistics* **5**, pp. 143–154.

H.-G. Müller (1984). Smooth optimum kernel estimators of densities regression curves and modes, *Annals of Statistics*, in press.

B. L. S. Prakasa Rao (1969). Estimation of a unimodal density, *Sankhya Series A* **31**, pp. 23–36.

Yu. V. Prohorov (1949). On the strong law of large numbers (in Russian), *Dokl. Akad. Nauk USSR* **69**.

M. Quenouille (1956). Notes on bias in estimation, *Biometrika* **43**, pp. 353–360.

J. W. Raatgever and R. P. W. Duin (1978). On the variable kernel model for multivariate nonparametric density estimation, in *COMPSTAT 78*, L. C. A. Corsten and J. Hermans (Eds.), Physica Verlag, Wien.

L. Rejtö and P. Révész (1973). Density estimation and pattern classification, *Problems of Control and Information Theory* **2**, pp. 67–80.

T. Robertson (1967). On estimating a density which is measurable with respect to a σ-lattice, *Annals of mathematical Statistics* **38**, pp. 482–493.

M. Rosenblatt (1971). Curve estimates, *Annals of Mathematical Statistics* **42**, pp. 1815–1842.

W. R. Schucany, H. L. Gray, and D. B. Owen (1971). On bias reduction in estimation, *Journal of the American Statistical Association* **66**, pp. 524–533.

W. R. Schucany and J. P. Sommers (1977). Improvement of kernel type density estimators, *Journal of the American Statistical Association* **72**, pp. 420–423.

D. W. Scott, R. A. Tapia, and J. R. Thompson (1980). Nonparametric probability density estimation by discrete maximum penalized likelihood criteria, *Annals of Statistics* **8**, pp. 820–832.

B. W. Silverman (1982). On the estimation of a probability density function by the maximum penalized likelihood method, *Annals of Statistics* **10**, pp. 795–810.

J. P. Sommers (1972). Improved density estimation, Technical Report 114, Department of Statistics, Southern Methodist University, Dallas, Texas, 1972.

R. A. Tapia and J. R. Thompson (1978). *Nonparametric Probability Density Estimation*, The Johns Hopkins University Press, Baltimore.

G. R. Terrell and D. W. Scott (1980). On improving convergence rates for nonnegative kernel density estimators, *Annals of Statistics* **8**, pp. 1160–1163.

J. Van Ryzin (1970). On a histogram method of density estimation, Technical Report 226, Statistics Department, University of Wisconsin, Madison, Wisconsin.

J. Van Ryzin (1973). A histogram method of density estimation, *Communications in Statistics* **2**, pp. 493–506.

G. Wahba (1971). A polynomial algorithm for density estimation, *Annals of Mathematical Statistics* **42**, pp. 1870–1886.

G. Wahba (1975). Optimal convergence properties of variable knot, kernel and orthogonal series methods for density estimation, *Annals of Statistics* **3**, pp. 15–29.

G. Wahba (1976). Histosplines with knots which are order statistics, *Journal of the Royal Statistical Society B* **38**, pp. 140–151.

E. J. Wegman (1969). A note on estimating a unimodal density, *Annals of Mathematical Statistics* **40**, pp. 1661–1667.

E. J. Wegman (1970a). Maximum likelihood estimation of a unimodal density function, *Annals of Mathematical Statistics* **41**, pp. 457–471.

E. J. Wegman (1970b). Maximum likelihood estimation of a unimodal density, II, *Annals of Mathematical Statistics* **41**, pp. 2160–2174.

E. J. Wegman (1975). Maximum likelihood estimation of a probability density function, *Sankhya Series A* **37**, pp. 211–224.

E. J. Wegman and H. I. Davies (1979). Remarks on some recursive estimators of a probability density, *Annals of Statistics* **7**, pp. 316–327.

C. T. Wolverton and T. J. Wagner (1969a). Asymptotically optimal discriminant functions for pattern classification, *IEEE Transactions on Information Theory* **IT-15**, pp. 258–265.

C. T. Wolverton and T. J. Wagner (1969b). Recursive estimates of probability densities, *IEEE Transactions on Systems, Science and Cybernetics* **5**, p. 307.

H. Yamato (1971). Sequential estimation of a continuous probability density function and the mode, *Bulletin of Mathematical Statistics* **14**, pp. 1–12.

CHAPTER 8

Simulation, Inequalities, and Random Variate Generation

1. CHOOSING A CRITERION

Consider the situation where one needs random variates with distribution function F on R^d, but uses random variates with distribution function G instead. The reasons for this replacement are sometimes economical (random variates from G are obtainable in less time or with less space) and sometimes practical (for the particular application a good approximation of F is all that is needed). Sometimes F is unknown and must be estimated from the data. And in many cases, one just does not want to spend a lot of time writing a complicated program for the generation of random variates with distribution function F. Whatever the reason for the replacement may be, it is necessary to have a good understanding of its consequences. How should one measure the goodness of the approximation for simulation purposes?

One of the classical criteria,

$$\Delta_1 = \sup_x |F(x) - G(x)|,$$

has the disadvantage that it is not sensitive to local discrepancies between the distributions. For example, if F puts all its mass uniformly on $[0, 1]$, $[2, 3], \ldots, [2n - 2, 2n - 1]$, and G puts all its mass uniformly on $[1, 2]$, $[3, 4], \ldots, [2n - 1, 2n]$, then $\Delta_1 = 1/n$. For large n, this is quite small, although it is clear that one would be reluctant to replace F by G in any simulation.

Assume that $d = 1$. If F and G are continuous and U is a uniform $[0, 1]$ random variable, then $F^{-1}(U)$ and $G^{-1}(U)$ are random variables with distribution functions F and G, respectively. This fact is of course at the basis of the *inversion method* in random variate generation. This leads to the

criterion

$$\Delta_2 = \sup_{0<u<1} |F^{-1}(u) - G^{-1}(u)|.$$

Unfortunately, Δ_2, like Δ_1, is not locally sensitive, and Δ_2 overemphasizes the tails of the distributions. For example, if F has infinite support and G has compact support, then $\Delta_2 = \infty$.

Consider now the *total variation criterion*

$$J = \frac{1}{2}\int |f - g| = \sup_B \left| \int_B f - \int_B g \right|,$$

where f and g are the densities corresponding to F and G. As explained in Chapter 1, J is an absolute bound on the error committed by replacing any probability $\int_A f$ by its approximation, $\int_A g$. If random variates are required for the purpose of the Monte Carlo evaluation of a functional $\int h\,dF$ (with $h \geq 0$), then

$$\left| \int h\,dF - \int h\,dG \right| = \left| \int_0^\infty \int_{h(x)\geq t} dF(x)\,dt - \int_0^\infty \int_{h(x)\geq t} dG(x)\,dt \right|$$

$$\leq \int_0^\infty \left| \int_{h(x)\geq t} dF(x) - \int_{h(x)\geq t} dG(x) \right| dt \leq J \sup_x h(x).$$

Hence, for bounded functions h, we have a clear upper bound on the error committed if a perfect evaluation of $\int h\,dG$ were possible. Often, J can be determined without much effort, but in some cases it is very hard to compute. In Section 2, we give several inequalities that may help in the determination of upper bounds for J.

2. INEQUALITIES

In this section, we give inequalities that link $\int |f - g|$ to other measures of the distance between f and g. Several of these inequalities were used in previous chapters in proofs of convergence. Other inequalities are helpful because they allow the user to infer some property about another distance measure from L_1 properties.

We start with inequalities that are useful in random variate generation.

THEOREM 1.

$$\int |f - g| \leq 2\min(K_r, K_c),$$

where

$$K_r = \sup\left(1 - \frac{g}{f}\right), \qquad K_c = \sup\left(1 - \frac{f}{g}\right).$$

Proof. By Theorem 1.1,

$$\int |f - g| = 2\int_{f \geq g} (f - g) = 2\int_{f \geq g} f\left(1 - \frac{g}{f}\right) \leq 2K_r.$$

Theorem 1 now follows by symmetry.

K_r and K_c are constants that are related to the rejection and composition methods in random variate generation. For example, the rejection algorithm uses K_r as follows:

Step 1. Generate a random variate X with density g, and an independent uniform $[0, 1]$ random variate U.

Step 2. *If* $Ug(X) \leq (1 - K_r)f(X)$ *then* exit with X *else* go to 1.

In the composition method, we proceed as follows:

Step 1. Generate a uniform $[0, 1]$ random variate U.

Step 2. *If* $U \leq 1 - K_c$ *then* exit with a random variate X having density g *else* exit with a random variate X having density $(f - (1 - K_c)g)/K_c$.

The random variates generated by both algorithms have density f provided we can do arithmetic with infinite precision. Both algorithms exploit the fact that g is a density that is close to f (i.e., has a small value for K_c or K_r), and that random variate generation for g is simple.

In Chapter 4, we needed an information theoretic inequality between the L_1 error and $\int f \log(f/g)$.

THEOREM 2.

$$\int f \log\left(\frac{f}{g}\right) \geq \frac{1}{2}\left(\int |f - g|\right)^2,$$

$$\frac{1}{2}\int |f - g| \leq \sqrt{1 - \exp\left(-\int f \log\left(\frac{f}{g}\right)\right)},$$

and

$$\int \min(f, g) \geq \frac{1}{2}\exp\left(-\int f \log\left(\frac{f}{g}\right)\right).$$

Proof. Let $A = \{f \geq g\}$, $B = \{f < g\}$, $h = gI_A/\int_A g$. Then, by Jensen's inequality,

$$\begin{aligned}\int_A f \log\left(\frac{f}{g}\right) &= \int h \frac{f}{g} \log\left(\frac{f}{g}\right) \int_A g \\ &\geq \int h \frac{f}{g} \log\left(\int h \frac{f}{g}\right) \int_A g \\ &= \int_A f \log\left(\frac{\int_A f}{\int_A g}\right).\end{aligned}$$

Define $p = \int_A f$, $q = \int_A g$. We have, by symmetry,

$$\int f \log\left(\frac{f}{g}\right) \geq p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right) = H(p, q).$$

Assume without loss of generality that $p = q + r$ for some $r > 0$. Write $H(p, q)$ as $H(q, r) = (q + r)\log(1 + r/q) + (1 - q - r)\log(1 - r/(1 - q))$, and note that $H'(q, r) = \log(1 + r/q) - \log(1 - r/(1 - q))$, and that $H''(q, r) = 1/p(1 - p) \geq 4$, where all the derivatives are with respect to r. Thus, by Taylor's expansion with remainder term,

$$H(p, q) \geq 4\left(\frac{r^2}{2}\right) = 2r^2 = 2\left(\int_{f>g} (f - g)\right)^2 = \frac{1}{2}\left(\int |f - g|\right)^2.$$

This concludes the proof of the first inequality.

For the second inequality, we employ Jensen's inequality once more:

$$\begin{aligned}-\int f \log\left(\frac{f}{g}\right) &= \int f \left(\log\left(\min\left(\frac{g}{f}, 1\right)\right) + \log\left(\max\left(\frac{g}{f}, 1\right)\right)\right) \\ &\leq \log\left(\int \min(f, g)\right) + \log\left(\int \max(f, g)\right).\end{aligned}$$

Thus,

$$\exp\left(-\int f \log\left(\frac{f}{g}\right)\right) \le \int \min(f,g) \int \max(f,g)$$

$$= \left(1 - \frac{1}{2}\int |f-g|\right)\left(1 + \frac{1}{2}\int |f-g|\right)$$

$$= 1 - \left(\frac{1}{2}\int |f-g|\right)^2.$$

The third inequality follows trivially from the previous argument.

The first inequality of Theorem 2 was proved by Kullback (1967), Csiszar (1967), and Kemperman (1969). The other inequalities and their proofs are due to Bretagnolle and Huber (1979). There is another distance measure that is closely related to $\int f \log(f/g)$ in the sense that both are finite or infinite simultaneously: $\int f^2/g - 1 = \int (f^2 - g^2)/g$. Its connection with the previous distance measures is given in Theorem 3.

THEOREM 3.

$$\int \frac{f^2}{g} - 1 \ge \int f \log\left(\frac{f}{g}\right) \ge \log\left(\int \frac{f^2}{g}\right),$$

and

$$\int \frac{f^2}{g} - 1 \ge \left(\int |f-g|\right)^2.$$

Also, for any function g, not necessarily a density,

$$\int \sqrt{f}\,|f-g| \le \sqrt{\int (f-g)^2}\,.$$

Proof. The left-hand side of the first inequality follows from the observation that $\log u \le u - 1$, all $u > 0$. The right-hand side follows directly from Jensen's inequality. The second inequality can be obtained by applying Hölder's inequality. Let $p, q > 1$ be such that $1/p + 1/q = 1$. Then,

$$\int |f-g| = \int \frac{|f-g|}{f^{1/p}} f^{1/p} \le \left(\int \frac{|f-g|^q}{f^{q/p}}\right)^{1/q} \left(\int f\right)^{1/p}.$$

The inequality follows after setting $p = q = 2$.

The last statement follows directly from the Cauchy–Schwarz inequality.

The Hellinger distance $H_p = (\int |f^{1/p} - g^{1/p}|^p)^{1/p}$, $p \geq 1$, shares with the L_1 error H_1 a few nice properties: it is always finite, and remains invariant under strictly monotone transformations. H_2 was suggested by Pitman (1979) as an aid in the study of maximum likelihood estimation. Unfortunately, there is no linear relationship between H_2 and $H_1 = \int |f - g|$, that is, there does not exist a universal constant a such that $H_2 \sim aH_1$ as $H_1 \to 0$. In fact, we have the following inequalities:

THEOREM 4.

$$H_2^2 \leq H_1 \leq H_2\sqrt{4 - H_2^2} \leq 2H_2.$$

Also, for any f, there exist sequences of densities f_n and g_n such that

$$H_1(f, f_n) \sim 2H_2^2(f, f_n) \to 0,$$

$$H_1(f, g_n) \sim 2H_2(f, g_n) \to 0.$$

Proof.

$$H_1 = \int |f - g| = \int |\sqrt{f} - \sqrt{g}|(\sqrt{f} + \sqrt{g}) \geq \int |\sqrt{f} - \sqrt{g}|^2 = H_2^2,$$

and

$$H_1^2 \leq \int (\sqrt{f} - \sqrt{g})^2 \int (\sqrt{f} + \sqrt{g})^2 = H_2^2\left(2 + 2\int \sqrt{fg}\right) = H_2^2(4 - H_2^2),$$

where we used the Cauchy–Schwarz inequality.

The sequence g_n is constructed by using a lot of overlap between f and g_n. Let m be a median for f. Set $g_n = (1 + p_n)f$ on $(-\infty, m]$ and $g_n = (1 - p_n)f$ on (m, ∞) for some sequence $p_n \downarrow 0$. Clearly, g_n is a density for each n. Also, $H_1 = p_n$, and $H_2^2 = 2 - 2\int \sqrt{fg_n} = 2 - \sqrt{1 - p_n} - \sqrt{1 + p_n} \sim p_n^2/4$. Here we used the fact that $\sqrt{1 - x} = 1 - x/2 - x^2/8 + O(x^3)$ as $x \downarrow 0$, and that $\sqrt{1 + x} = 1 + x/2 - x^2/8 + O(x^3)$ as $x \downarrow 0$.

The equality $H_1 = H_2^2$ is attained for nonoverlapping f and g, that is, $\int \sqrt{fg} = 0$. The sequence f_n is therefore partially based upon a sequence of densities that is nonoverlapping with f. Let m_n be the $1/n$ quantile of f, and define f_n by: $f_n = \sqrt{n}f$ on $(-\infty, m_n]$, $f_n = (1 - 1/\sqrt{n})f/(1 - 1/n)$ on (m_n, ∞). Again f_n is a density for each n. We verify easily that $H_1 = 2(\sqrt{n} - 1)/n \sim 2/\sqrt{n}$. Also, $H_2^2 = 2 - 2\int \sqrt{ff_n} = 2 - 2/n^{3/4} - 2\sqrt{(1 - 1/n)(1 - 1/\sqrt{n})} \sim 2 - 2n^{-3/4} - 2(1 - 1/2/\sqrt{n}) \sim 1/\sqrt{n}$.

In Theorem 5, we state *LeCam's inequality* (1973), which was used in a crucial place in the proof of Assouad's Lemma (Theorem 4.5).

THEOREM 5. *For any densities f and g on R^d,*

$$\int \min(f, g) \geq \tfrac{1}{2}\left(\int \sqrt{fg}\right)^2.$$

Proof. By the Cauchy–Schwarz inequality

$$\left(\int_{f<g} \sqrt{fg}\right)^2 = \left(\int_{f<g} f\sqrt{\frac{g}{f}}\right)^2 \leq \int_{f<g} f\frac{g}{f} \int_{f<g} f \leq \int_{f<g} f.$$

By symmetry,

$$\left(\int \sqrt{fg}\right)^2 \leq 2\int_{f<g} f + 2\int_{g\leq f} g = 2\int \min(f, g).$$

We finally consider the sup norm $\operatorname{ess\,sup}|f - g|$, where the essential supremum is with respect to Lebesgue measure. It is clear that $\int|f - g|$ can be small while $\sup|f - g|$ is large, possibly infinite. Vice versa, a small sup norm does not guarantee a small L_1 distance, unless one is willing to make assumptions about the tail of f or g. As an example of how this can be done, we cite a few inequalities due to Serfling (1979).

THEOREM 6. *Let f and g be densities in R^d, and let r be a positive constant. Let v_d be $\pi^{d/2}\Gamma(d/2 + 1)$, and define*

$$A = \begin{cases} \sup_t t^r \int_{\|x\|>t} f\left(\leq \int \|x\|^r f\right), & 0 < r < \infty; \\ \inf\left\{t: \int_{\|x\|\leq t} f = 1\right\}, & r = \infty. \end{cases}$$

Then

$$\int|f - g| \leq 4A^{d/(r+d)}\left(v_d \operatorname{ess\,sup}|f - g|\right)^{r/(r+d)}, \qquad 0 < r < \infty.$$

and

$$\int|f - g| \leq 2A^d v_d \operatorname{ess\,sup}|f - g|, \qquad r = \infty.$$

In the definition of A, we can replace f by g if we wish.

Proof.

$$\frac{1}{2}\int|f-g| = \int_{f>g}(f-g) = \int_{\substack{\|x\|\le t\\ f>g}}(f-g) + \int_{\substack{\|x\|>t\\ f>g}}(f-g)$$

$$\le v_d t^d \operatorname{ess\,sup}|f-g| + At^{-r}, \quad \text{all } t>0.$$

The terms on the right-hand side are equal for $t^{r+d} = A/(v_d \operatorname{ess\,sup}|f-g|)$. Resubstitution gives the first inequality. The second inequality is straightforward.

3. THE GENERALIZATION OF A SAMPLE FOR RANDOM VARIATE GENERATION

We are given a sample $X_1, \ldots, X_n$ of independent R^d-valued random vectors with common unknown density f, and are asked to generate (i.e., produce by means of a computer) a new independent sample $Y_1, \ldots, Y_m$ of independent random vectors with the same density f. Stated in this manner, the problem has obviously no solution. Some of the obstacles can be bypassed by convenient and not so unrealistic assumptions:

(i) Real numbers can be stored on a computer; otherwise, the notion of "density" would be vacuous.

(ii) We have a source capable of generating a sequence $U_1, U_2, \ldots$ of independent random variables uniformly distributed on $[0,1]$.

Since f is unknown, it must either explicitly or implicitly be estimated from $X_1, \ldots, X_n$. In all generality, we are interested in procedures that take the following format:

Step 1. Construct a density estimate $f_n(x) = f_n(x; X_1, \ldots, X_n)$ of $f(x)$.

Step 2. *For* $i = 1$ *to* m *do*:
Generate a new uniform $[0,1]$ random variable U_i.
Compute Y_i from f_n and U_i.

It is clear that both samples are dependent. Also, unless we are incredibly lucky, f_n is not equal to f. In this section we will discuss to what extent these undesirable effects can be limited.

The following topics are of particular interest to us: sample independence; consistency; sample indistinguishability; moment matching; generators for f_n.

3.1. Sample Independence

There is very little that can be done about the dependence between $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ except to hope that for n large enough, some sort of asymptotical independence is approached. Also, in some applications, sample independence is not a requirement at all.

Since $Y_1, \ldots, Y_m$ are independent given $X_1, \ldots, X_n$, we need only consider the dependence between $Y = Y_1$ and $X_1, \ldots, X_n$. A measure of this dependence is

$$D_n = \sup_{A,B} |P(Y \in A, X \in B) - P(Y \in A)P(X \in B)|,$$

where the supremum is with respect to all Borel sets A of R^d and all Borel sets B of R^{nd}, and where X is our short notation for $(X_1, \ldots, X_n)$. We say that the samples are asymptotically independent when

$$\lim_{n \to \infty} D_n = 0.$$

In situations where $X_1, \ldots, X_n$ is used to design or build a system, and $Y_1, \ldots, Y_m$ is used to test it, the sample dependence will often cause optimistic evaluations. Without the asymptotical independence, we can't even hope to diminish this optimistic bias by increasing n.

The inequality of Theorem 7 below provides us with a sufficient condition for asymptotical independence: $\lim_{n \to \infty} E(J_n) = 0$.

THEOREM 7.

$$D_n \le E(J_n) = E\left(\int |f_n - f|\right).$$

Proof. We have

$$\begin{aligned} D_n \le & \sup_{A,B} |P(Y \in A, X \in B) - P(X_{n+1} \in A, X \in B)| \\ & + \sup_{A,B} |P(X_{n+1} \in A, X \in B) - P(X_{n+1} \in A)P(X \in B)| \\ & + \sup_{A,B} |P(X_{n+1} \in A)P(X \in B) - P(Y \in A)P(X \in B)|. \end{aligned} \quad (1)$$

The last term of (1) is equal to

$$\sup_A |P(X_{n+1} \in A) - P(Y \in A)| = \sup_A \left| \int_A E(f_n) - \int_A f \right|$$

$$= \frac{1}{2} \int |E(f_n) - f|$$

by Scheffé's Theorem 1.1. The second term of (1) is obviously 0, while the first term does not exceed

$$\sup_{A,B} E\left(I_{[X \in B]} \left| \int_A f_n - \int_A f \right| \right) \leq \sup_A E\left(\left| \int_A f_n - \int_A f \right| \right)$$

$$\leq E\left(\sup_A \left| \int_A f_n - \int_A f \right| \right).$$

$$= E\left(\frac{1}{2} \int |f_n - f| \right).$$

This concludes the proof of Theorem 7.

3.2. Consistency

Theorem 7 shows that asymptotical independence of the samples follows from consistency of the density estimate, that is, $\lim_{n \to \infty} E(J_n) = 0$. But more importantly, consistency is needed for good approximations of all the probabilities because

$$\sup_A \left| \int_A f_n - \int_A f \right| = \frac{1}{2} \int |f_n - f| = \frac{1}{2} J_n. \tag{2}$$

(see Theorem 1.1, the discussion of Chapter 1, and Section 1 of this chapter).

3.3. Sample Indistinguishability

In simulations, one important measure of the goodness of a method is the indistinguishability of $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_m$ *for the given sample size* m. When Card(A) and Card*(A) are the cardinalities of A for $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_m$, respectively, where A is an arbitrary set of R^d, then this

could be measured by

$$S_n = \sup_A |E(\text{Card}(A)) - E(\text{Card}^*(A)|X_1, \ldots, X_n)|.$$

When the random variable S_n is smaller than 1, then *all* sets A capture in each sample on average between $u - \frac{1}{2}$ and $u + \frac{1}{2}$ points for some u. Such a strict criterion is needed, for example, when extremal sets become important.

But a little thought shows that

$$S_n = m \sup_A \left| \int_A f_n - \int_A f \right| = \frac{m}{2} \int |f_n - f| = \frac{m}{2} J_n.$$

Thus once again, we are led to the L_1 criterion J_n.

We say that f_n is *k-excellent for samples of size m* when

$$E(S_n) = \frac{m}{2} E(J_n) \leq k. \tag{3}$$

The notion of 1-excellence is very strong. To illustrate this, we will show just how poorly any parametric or nonparametric density estimate must perform. In Table 1 we have calculated various threshold values for n below which we cannot have 1-excellence for fixed m, and this for certain combinations of density estimates f_n and densities f. The following combinations are considered:

A. *All* estimates f_n, and some f of the form $p\tilde{f} + (1 - p)g$, where $\tilde{f}$ and g are known densities with disjoint supports, and p is the only unknown, a number between 0 and 1.

B. *All* estimates f_n, and some $f \in F_{2,r}$, where r is arbitrary but at least equal to $2r^* = (3888)^{1/5} = 5.2233033 \cdots$ (see Chapter 4). This is essentially the class of all densities with bounded values of $B^*(f)$.

TABLE 1

m	A	B	C	D
10	1	1	40	85
100	18	1	13,000	85,000
1,000	1,800	15	4,000,000	85,000,000
10,000	180,000	4,900	1,300,000,000	85,000,000,000
100,000	18,000,000	1,500,000	400,000,000,000	85,000,000,000,000

C. *All* kernel estimates (all K and h are allowed), and *all* f.

D. *All* histogram estimates and all $f \in \mathscr{F}$, that is, all absolutely continuous f with bounded and continuous a.e. derivative f'.

The figures of Table 1 were obtained from results of Chapters 4 and 5 after leaving out the $o(1)$ terms in all asymptotic expansions, and rounding to two decimal digits. They should therefore only be considered as approximate figures. For A and B, we relied on the approximations provided by Theorem 4.4

$$E(J_n) \geq \frac{0.0849856\cdots}{\sqrt{n}} \qquad \left(\text{or } n \geq (0.0424928\cdots m)^2\right) \tag{4}$$

and Theorem 4.3

$$E(J_n) \geq (2e)^{-4}(3888)^{1/5}n^{-2/5} \qquad \left(\text{or } n \geq (0.00298963\cdots m)^{5/2}\right), \tag{5}$$

respectively. The lower figures of column B can be explained by the sloppiness of the argument of Theorem 4.4. Also, if $2r^*$ is replaced by $8r^*$, the threshold value for n increases by a factor of $4^{5/2} = 32$. For $m = 10000$, this would imply that for any density estimate, however good, there is a density in $F_{2,r}, r \leq 8r^*$, for which we cannot have 1-excellence when $n < 32 \times 4900 = 156{,}800$. Even for the simple class A, where we can use very simple parametric estimates, for 1-excellence uniformly over all f in this class and $m = 10000$, at least $n = 180000$ original data points are needed. Columns A and B both tell us that there exists some density in a class of densities for which there cannot be any 1-excellence when n is too small. The particular density for which this happens is unknown: it depends upon the f_n that is used, and upon n. This is precisely the weakness of the lower bounds of Chapter 4: we are still not satisfied with this result because it is after all possible that for the f that we are trying to estimate, we have 1-excellence, despite Table 1. The individual bounds of Chapter 5 are more powerful in this respect.

For example, for any standard kernel estimate, and all f, we cannot possibly have 1-excellence for $m = 1{,}000$ when $n < 4{,}000{,}000$. This is a powerful statement, derived from Theorem 5.2:

$$E(J_n) \geq C_3 n^{-2/5} \qquad \left(\text{or } n \geq (0.43933402\cdots m)^{5/2}\right). \tag{6}$$

There is no exception: it applies to all densities. Thus, the 4 million lower bound on the sample size is an absolute lower bound. If the figures for the kernel estimate (column C) are disappointing, then the figures for the

histogram estimate are even more disappointing (column D): the lower bound for $m = 1{,}000$ becomes now 85 million, and the rate of increase in column D is as m^3. The figures are based on Theorem 5.5:

$$E(J_n) \geq 0.880261 \cdots n^{-1/3} \qquad \left(\text{or } n \geq (0.4401305 \cdots m)^3\right). \quad (7)$$

The user may sometimes wish to relax his requirement to k-excellence with $k > 1$, because the high quality that is inherent in the condition $E(S_n) \geq 1$ is often not needed. In that case, the m-column in Table 1 should be multiplied with k.

Nevertheless, Table 1 demonstrates very clearly that for larger values of m, we should not use a histogram estimate, unless there are other more important factors than $E(J_n)$ such as the monotonicity of the random variate generator (see Section 3.5 below).

The bounds (4)–(7) are negative results. On the positive side, it is reassuring to know that we can indeed achieve 1-excellence with the kernel estimate for any m if we are willing to pay a price for it. In Theorem 5.1 we have shown that in first approximation, for all f with compact support, if the optimal h and K are chosen,

$$E(J_n) \leq 1.3768102 \cdots \left(\frac{9}{125}\right)^{1/5} B^*(f) n^{-2/5}.$$

Another bound is provided by Theorem 5.10:

$$E(J_n) \leq 1.240701 \cdots B_H^*(f) n^{-1/3}.$$

These two approximate inequalities can be used to conclude that if f is isosceles triangular, 1-excellence is obtained whenever

$$n \geq \left(\frac{1.17624440 \cdots}{2} m\right)^{5/2},$$

and that if f is uniform $[0, 1]$, the same is true when

$$n \geq \left(\frac{1.240701 \cdots}{2} m\right)^3,$$

at least when h is optimally chosen, and K is the Epanechnikov kernel and the isosceles triangular kernel, respectively. The figures of Table 2 illustrate these inequalities. In case of k-excellence, multiply the m-column by k.

TABLE 2

m	Uniform $[0,1]$ f	Isosceles Triangular f
10	239	83.9
100	239,000	26,500
1,000	239,000,000	8,390,000
10,000	239,000,000,000	2,650,000,000
100,000	239,000,000,000,000	839,000,000,000

3.4. Moment Matching

Some statisticians and engineers attach great importance to the moments of the densities f_n and f. For $d = 1$, the ith moment mismatch is the following random variable (defined when $|x|^i$ is integrable with respect to f_n and f):

$$M_{ni} = \int x^i f_n - \int x^i f, \qquad i = 1, 2, 3, \ldots . \tag{8}$$

In Theorem 8, we give M_{n1} and M_{n2} for the kernel estimate:

THEOREM 8. *For the kernel estimate on* R^1, *with* $\int xK = 0$, $\int x^2 K = \sigma^2$,

$$M_{n1} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - E(X_i)\right), \qquad M_{n2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - E\left(X_i^2\right)\right) + h^2\sigma^2.$$

Also,

$$E(M_{n1}) = 0, \qquad \operatorname{Var}(M_{n1}) = \frac{\operatorname{Var}(X_1)}{n}, \qquad E(M_{n2}) = h^2\sigma^2$$

and

$$\operatorname{Var}(M_{n2}) = \frac{\operatorname{Var}\left(X_1^2\right)}{n}.$$

Proof. We use the fact that f_n is the density of the random variable $Y = X_Z + hW$, where Z, W are independent of the X_i's and of each other, Z is uniform on $\{1, \ldots, n\}$, and W has density K. Theorem 8 follows from the observation that

$$E(Y|X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i + hE(W) = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

and

$$E(Y^2|X_1,\ldots,X_n) = E(X_Z^2 + 2hWX_Z + h^2W^2|X_1,\ldots,X_n)$$

$$= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + h^2\sigma^2.$$

We note first that the distribution of M_{n1} is not influenced by the choice of h or K. By the weak law of large numbers, M_{n1} tends in probability to 0 when $E(|X_1|) < \infty$, regardless of how h varies with n. The second moment mismatch however consists of a random variable not influenced by h or K plus a constant $h^2\sigma^2$. This constant is the sole contributor to the positive bias in $E(M_{n2}) = h^2\sigma^2$. Since we have no control over $\mathrm{Var}(M_{n2})$, the best we can do is to make the bias as small as possible. But this would force us to choose h so small that $E(J_n)$ increases. If h is chosen optimally for $E(J_n)$, then it varies as $n^{-1/5}$ for many smooth distributions. In that case, $E(M_{n2})$ becomes a good measure of the second moment mismatch in view of $\sqrt{\mathrm{Var}(M_{n2})} = O(n^{-1/2}) = o(E(M_{n2}))$.

For example, when K is the Epanechnikov kernel and h is chosen optimally (see (5.14)), then the normalized second moment mismatch is

$$\frac{E(M_{n2})}{\mathrm{Var}(X_1)} = \frac{1}{5}\left(\frac{15}{2\pi}\right)^{2/5} \frac{n^{-2/5}\left(\int\sqrt{f}/\int|f''|\right)^{4/5}}{\mathrm{Var}(X_1)}. \tag{9}$$

If we had used the optimal h for a given K, then we would have obtained an expression attaining its minimal value for the Epanechnikov kernel. Thus, the choice of the Epanechnikov kernel is well motivated. Expression (9) is translation and scale invariant. Only the shape of f is important. To get a rough idea of the size of (9), we can take the normal density as our prototype. We obtain

$$\frac{E(M_{n2})}{\mathrm{Var}(X_1)} = \left(\frac{225\pi e^2}{32}\right)^{1/5}\frac{1}{5}n^{-2/5} = 0.5540591\cdots n^{-2/5}. \tag{10}$$

Table 3 gives different values of n (derived from (10)) needed to achieve specific percentage relative errors for the second moment mismatch for the normal density. For standard values of n, we note that this error ranges from 1 to 10%. Thus, the smoothing necessary for consistency and small values of $E(J_n)$ has an undesirable side-effect on the second moment mismatch, and this effect is especially outspoken for small n. We also note

TABLE 3

n	Normalized Second Moment Mismatch, Normal, f, Optimal h
10	$0.2205\cdots$
100	$0.08781\cdots$
1,000	$0.03496\cdots$
10,000	$0.01391\cdots$
100,000	$0.005540\cdots$
1,000,000	$0.002205\cdots$

from Table 3 that it is all but hopeless to ask for a relative error of the order of 0.1% or less. The situation for $d > 1$ is of course more complex (see, e.g., Shanmugam, 1977, for a related discussion).

We also observe that, by (5.18), the normalized second moment mismatch does not exceed

$$\frac{1}{5} \times (6.7726100\cdots)^2 n^{-2/5} = 9.1736492\cdots n^{-2/5}$$

when K is Epanechnikov's kernel and h is chosen as in (5.14). Various values for this universal bound are given in Table 4.

3.5. Generators for f_n

For the kernel estimate

$$f_n(x) = (nh^d)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

TABLE 4

n	Normalized Second Moment Mismatch, Absolute Upper Bound
10	$3.652\cdots$
100	$1.453\cdots$
1,000	$0.5788\cdots$
10,000	$0.2304\cdots$
100,000	$0.09173\cdots$
1,000,000	$0.03652\cdots$

the following procedure can be used for generating random variates:

Step 1. Generate Z uniformly on $\{1, \ldots, n\}$, and generate an independent random vector W with density K.

Step 2. Exit with $Y \leftarrow X_Z + hW$.

The only possible complication is created by the kernel K. For $d = 1$, we have seen several arguments for choosing the Epanechnikov kernel

$$K(x) = \tfrac{3}{4}(1 - x^2), \qquad |x| \le 1.$$

There are two very fast algorithms for generating a random variate W with this density—the rejection method and the order statistics method:

(The rejection method with a rectangular dominating density)

Step 1. *Repeat* Generate a uniform $[-1, 1]$ random variate W, and an independent uniform $[0, 1]$ random variate U.
Until $U \le 1 - W^2$.

Step 2. Exit with W.

(The order statistics method)

Step 1. Generate three independent uniform $[-1, 1]$ random variates V_1, V_2 and V_3. Set $W \leftarrow V_3$.

Step 2. If $|V_3| > |V_1|$ and $|V_3| > |V_2|$, set $W \leftarrow V_2$. Exit with W.

In the rejection method, W is accepted in Step 2 with probability $\frac{2}{3}$, so that, per W produced, three uniform $[-1, 1]$ random variates are used, on average. However, we also need some multiplications. The order statistics method also requires three uniform random variates, but the multiplication is replaced by a few absolute value operations.

In R^d, the optimal K according to L_2 criteria was computed by Deheuvels (1977b). It takes the form

$$K(x) = C_d\left(d + 4 - \|x\|_2^2\right), \qquad \|x\|_2^2 \le d + 4,$$

where C_d is a normalization constant depending upon d only. Random vectors with this density (the multivariate Pearson II density) can be obtained as

$$\sqrt{d+4}\,\sqrt{\text{Beta}(d/2, 2)}\; T_d,$$

where $\text{Beta}(d/2, 2)$ is a beta random variable, and T_d is an independent

random vector uniformly distributed on the unit sphere of R^d. For beta random variate generation we refer to the work of Schmeiser (Schmeiser and Shalaby, 1980; Schmeiser and Babu, 1980), and the references found in these papers and in Schmeiser (1980). T_d can be generated easily by a variety of methods, for example, the spacings method (Sibuya, 1962; Tashiro, 1977), the polar method, or specific methods for small values of d (see the survey papers by Deak, 1979, and Rubinstein, 1980).

For example, in the polar method one exploits the fact that T_d is distributed as $(N_1/N, \ldots, N_d/N)$, where $N_1, \ldots, N_d$ are independent normal $(0,1)$ random variables, and $N = \sqrt{N_1^2 + \cdots + N_d^2}$.

For the choice of h as a function of the data, we refer to Sections 5.6 and 6.2. In some contexts, further modifications of the kernel estimate may be needed, requiring some modifictions in the generation algorithm. For $d = 1$, $E(J_n)$ is usually reduced by using the transformed kernel estimate defined in Chapter 9. For $d > 1$, there is no equivalent of the transformed kernel estimate. Directional information in the data can be used to improve the performance of the kernel estimate. For example, Shanmugam (1977) (see also Deheuvels, 1977b) discusses what happens when instead of $Y = X_Z + hW$ we use $Y = X_Z + hAW$, where A is a $d \times d$ matrix chosen such that $A'A$ is equal to the inverse of the sample covariance matrix. The estimate of Breiman, Meisel, and Purcell (1977, see Chapter 7) requires modifying the basic algorithm to $Y = X_Z + h_Z W$ (and thus storing $h_1, \ldots, h_n$ together with $X_1, \ldots, X_n$), where h_i is the distance of X_i to its kth nearest neighbor among $X_1, \ldots, X_n$, and k is an integer to be selected beforehand.

Another modification is needed when f is known to concentrate all its mass on $[0, \infty)$ or $[0,1]$ (see Chapter 9 below, or Hominal and Deheuvels, 1979): f_n can be replaced by $\bar{f}_n = f_n / \int_0^\infty f_n$ on $[0, \infty)$ and by $\bar{f}_n = 0$ on $(-\infty, 0)$. As proved in Theorem 11.3, this replacement is totally harmless for any estimate. Random variate generation is no problem either:

Step 1. *Repeat* Generate X with density f_n *Until* $X \geq 0$.

Step 2. Exit with X.

The average number of executions of Step 1 in this rejection algorithm is $1/\int_0^\infty f_n$, which is usually close to 1.

The bias reduction devices of Section 7.6 cause only minor inconveniences. Consider first the normalized Bartlett estimate $g_n = (f_n)_+ / \int (f_n)_+$, where f_n is a kernel estimate with kernel K taking negative and positive values. It suffices to note that $(f_n)_+ \leq f_n^*$, where

$$f_n^*(x) = (nh)^{-1} \sum_{i=1}^{n} K_+\left(\frac{x - X_i}{h}\right).$$

We can now proceed by von Neumann's rejection method (von Neumann, 1951; see also, e.g., Rubinstein, 1981):

Step 1. *Repeat* Generate three independent random variates I, W, and U, where I is a random integer between 1 and n. W has density $K_+/\int K_+$, and U is uniform $[0,1]$. Set $X \leftarrow X_I + hW$. X now has density $f_n^*/\int f_n^*$.

Until $U \sum_{i=1}^n K_+((X - X_i)/h) \le (\sum_{i=1}^n K((X - Xi)/h))_+$.

Step 2. Exit with X.

The expected number of iterations of the repeat loop is equal to $\int f_n^*/\int (f_n)_+ = \int K_+/\int (f_n)_+ \le \int K_+$. The upper bound is tight because $\int (f_n)_+ \to 1$ when $h \to 0$, $nh \to \infty$ (this is a corollary of Theorem 7.4). For example, with Bartlett's kernel

$$K(x) = \frac{9}{8}\left(1 - \frac{5x^2}{3}\right), \qquad |x| \le 1,$$

we obtain

$$\int K_+ = 2\int_0^{\sqrt{3/5}} \frac{9}{8}\left(1 - 5\frac{x^2}{3}\right) dx = \sqrt{\frac{27}{20}} = 1.161895\cdots.$$

A more time-consuming operation however is the evaluation of the sums in the "until" statement. Obviously, one should never take the sums blindly, as this would cause the evaluation time to increase linearly with n. When K vanishes outside $[-1,1]$, as is the case in our example, we can, for example, store all the data in increasing order in an array. By binary search (see Knuth, 1975), we can determine in what interval (X_i, X_{i+1}) X falls. By searching up and down the array from the given interval, we can find all the X_i's that are within distance h of X. No other X_i's can possibly influence our sums. Under some conditions on f, one can show that the expected time now becomes $O(nh) + O(\log n)$.

Further reductions in evaluation time are possible if we take the form of K also into account. For example, when K is quadratic on $[-1,1]$ (examples include Epanechnikov's kernel and Bartlett's kernel), we know that f_n, the kernel estimate, is a piecewise quadratic spline function with breakpoints at $X_i - h$, $X_i + h$, $1 \le i \le n$. Thus, we need only store these breakpoints in order, together with for each interval the three coefficients of the quadratic polynomial. Once the interval to which X belongs is determined (by binary search, this can be done in time $O(\log n)$), the evaluation of the sums in the "until" statement becomes a snap (time $O(1)$). Admittedly, the

set-up time has gone up dramatically: it too can be kept within reasonable bounds by first sorting the data, and then computing all the coefficients in one extra pass of the data, from left to right. For uniform kernels K, f_n is piecewise constant, so that the previous procedure can be simplified somewhat. For what follows, we can and do assume that the kernel estimate can be evaluated in time $O(\log n)$ when preprocessing time is ignored.

For the estimate of Terrell and Scott, also presented in Section 7.6, we have an unnormalized estimate $f_n = f_{n1}(f_{n1}/f_{n2})^{1/3}$, where the correction factor $(f_{n1}/f_{n2})^{1/3}$ always takes values between 0 and $2^{1/3}$. Thus, we have $f_n \le 2^{1/3} f_{n1}$. In the algorithm for Bartlett's estimate, we must make only a few modifications: in Step 1, X has density f_{n1} and W has density K; and we iterate until $U2^{1/3} \le (f_{n1}/f_{n2})^{1/3}$. The expected number of iterations is $2^{1/3} + o(1)$ (a corollary of $\int f_n \to 1$; see Theorem 7.6), and the evaluation of $(f_{n1}(X)/f_{n2}(X))^{1/3}$ takes time $O(\log n)$, uniformly over all values for X.

Consider now the histogram estimate defined by partitions $\mathscr{P}_n = \{A_{nj}, j \text{ integer}\}$. An algorithm for generating random variates with density f_n is easy to find, for example:

Step 0. Preprocessing Compute and store the probabilities $p_i = \mu_n(A_{ni})$ for which $p_i \neq 0$ (μ_n is the standard empirical measure for $X_1, \ldots, X_n$). Note that not more than n p_i's need to be stored.

Step 1. Generate a random integer I such that $P(I = i) = p_i$.

Step 2. Generate a random variate X uniformly in A_{nI}, and exit.

In contrast to the algorithm given above for the kernel estimate, a preprocessing step is needed here. With some careful programming, it can be implemented in time $O(n \log n)$ because for each X_i we must check if the set A_{nj} it belongs to contains another X_m. An equivalent algorithm without the preprocessing step is the following:

Step 1. Generate Z uniformly on $\{1, \ldots, n\}$, and find the set A_{nj} to which X_Z belongs.

Step 2. Exit with Y uniformly distributed in A_{nj}.

This algorithm is preferable over the algorithm with preprocessing in all but a few special cases. One such special case occurs when maximally anticorrelated random variables are needed, for example, in variance reduction for Monte Carlo simulations. It is known that for a continuous distribution function F, two random variables with this distribution function and maximal anticorrelation can be obtained by using $F^{-1}(U)$ and $F^{-1}(1 - U)$, where U is a uniform $[0, 1]$ random variable. Fox (1980) argues that we should try to implement the inversion method (i.e., generate a random

variable with distribution function F as $F^{-1}(U)$) by all means. Unfortunately, none of the algorithms given above for the kernel and histogram estimates are based upon inversion.

We will conclude this section by discussing various inversion algorithms for density estimates. Most of the interesting estimates are piecewise polynomial: these include the histogram estimate, and the standard kernel estimate with uniform kernel, triangular kernel, or quadratic kernel vanishing outside $[-1, 1]$. Assume that the real line is partitioned by the breakpoints $a_1 < a_2 < \cdots < a_n$, and that the estimate vanishes outside $[a_1, a_n]$. On $[a_i, a_{i+1})$, the estimate takes the form

$$f_n(x) = b_{i0} + b_{i1}x + b_{i2}x^2 + \cdots + b_{ip}x^p.$$

Assume also that we know the value of the corresponding distribution function, F_n, at these breakpoints: $F_n(a_i) = c_i$. It is not difficult to verify that these coefficients, breakpoints, and values are indeed easy to compute for our estimates.

The inversion algorithm proceeds as follows:

Step 1. Generate a uniform $[0, 1]$ random variate U. Determine the integer I with the property that $c_I \le U < c_{I+1}$. (Thus, I takes values between 1 and $n - 1$, since $c_1 = 0$, and $c_n = 1$.)

Step 2. Exit with X, where X is the solution of the equation

$$U - c_I = b_{I0}(X - a_I) + \frac{1}{2}b_{I1}(X^2 - a_I^2) + \cdots + \frac{1}{p+1}b_{Ip}(X^{p+1} - a_I^{p+1}).$$

For a piecewise constant estimate, the solution of the equation is very simple. For piecewise quadratic estimates, the solution involves finding the roots of a polynomial of degree three. In any case, the time taken in Step 2 is $O(1)$, that is, it does not depend upon n. If a sequential interval search is employed in Step 1, its time could grow linearly with n. Also, binary interval search is not recommended because the time required increases as $\log n$ in the worst case. The alias method (Walker, 1977; Kronmal and Peterson, 1979) can be used to generate a random integer I distributed as our I of Step 1, in time $O(1)$ in the worst case. Unfortunately, it is ineligible because the integer is not obtained by inverting U. For Step 1, the prime candidate seems to be the method of "guide tables" (Chen and Asau, 1974; see also the comprehensive survey paper of Ahrens and Kohrt, 1981). It takes expected time $O(1)$, but could do much worse in the worst case.

The principle is very simple: in a preprocessing step, a guide table g_i, $1 \le i \le n$, is constructed, where

$$g_i = \max(j: c_j < i/n).$$

For example, it is clear that $g_n = n - 1$. The determination of I such that $c_I \le U < c_{I+1}$ is done by sequential search from a place determined by the guide table. Thus, the guide table is "almost" an inversion table.

Step 1. $I \leftarrow \underline{nU + 1}$. ($I$ now has the property $(I - 1)/n \le U < I/n$.)
Step 2. $I \leftarrow g_{I-1}$. (Look-up in the guide table.)
Step 3. *While* $c_{I+1} < U$ *do* $I \leftarrow I + 1$.
Step 4. Exit with I.

The interesting fact about this algorithm is that on the average the "while" loop is executed at most once. This is because n values of c_i are distributed over n intervals and U is uniform $[0, 1]$. The guide table itself can be constructed in linear time:

Step 1. *For* $i = 1$ *to* n *do*: $g_i \leftarrow 0$.
Step 2. *For* $j = 1$ *to* n *do*: $i \leftarrow \underline{nc_j + 1}$, $g_i \leftarrow j$. (Note: $(i - 1)/n \le c_j < i/n$.)
Step 3. *For* $i = 2$ *to* n *do*: $g_i \leftarrow \max(g_{i-1}, g_i)$. (Adjustment for empty intervals.)

We should observe that for histogram estimates with data-dependent breakpoints (i.e., breakpoints at the kth, $2k$th, etc. order statistics of the data), the implementation of the inversion method becomes extremely simple (Fox, 1980). Archer (1980), also concerned with simple generators for f_n, proposes a piecewise constant density estimate with breakpoints and heights determined in such a way that the moments of f_n match those of the data. Unfortunately, such an approach does not yield a consistent estimate in general. Thompson and Taylor (1982) report a method for generating random variates in R^d without explicitly constructing f_n.

REFERENCES

J. H. Ahrens and K. D. Kohrt (1981). Computer methods for efficient sampling from largely arbitrary statistical distributions, *Computing* **26**, pp. 19–31.

N. P. Archer (1980). The generation of piecewise linear approximations of probability distribution functions, *Journal of Statistical Computation and Simulation* **11**, pp. 21–40.

L. Breiman, W. Meisel, and E. Purcell (1977). Variable kernel estimates of multivariate densities, *Technometrics* **19**, pp. 135–144.

J. Bretagnolle and C. Huber (1979). Estimation des densités: risque minimax, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **47**, pp. 119–137.

H. C. Chen and Y. Asau (1974). On generating random variates from an empirical distribution, *AIIE Transactions* **6**, pp. 163–166.

I. Csiszar (1967). Information-type measures of difference of probability distributions and indirect observations, *Studia Scientiarum Mathematicarum Hungarica* **2**, pp. 299–318.

I. Deak (1979). Comparison of methods for generating uniformly distributed random points in and on a hypersphere, *Problems of Control and Information Theory* **8**, pp. 105–113.

P. Deheuvels (1977a). Estimation non paramétrique de la densité par histogrammes généralisés, *Revue de Statistique Appliquée* **25**, pp. 5–42.

P. Deheuvels (1977b). Estimation non paramétrique de la densité par histogrammes généralisés, *Publications de l'ISUP* **22**, pp. 1–23.

L. Devroye (1982). A note on approximations in random variate generation, *Journal of Statistical Computation and Simulation* **14**, pp. 149–158.

B. L. Fox (1980). Monotonicity, extremal correlations, and synchronization: implications for nonuniform random numbers, Technical Report, Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Canada.

P. Hominal and P. Deheuvels (1979). Estimation non paramétrique de la densité compte-tenu d'informations sur le support, *Revue de Statistique Appliquée* **27**, pp. 47–68.

J. H. B. Kemperman (1969). On the optimum rate of transmitting information, *Probability and Information Theory*, Springer Lecture Notes in Mathematics 89, Springer-Verlag, Berlin, pp. 126–169.

D. E. Knuth (1975). *The Art of Computer Programming, Vol. 3: Sorting and Searching*, Addison-Wesley, Reading, Massachusetts.

R. A. Kronmal and A. V. Peterson (1979). On the alias method for generating random variables from a discrete distribution, *The American Statistician* **33**, pp. 214–218.

S. Kullback (1967). A lower bound for discrimination information in terms of variation, *IEEE Transactions on Information Theory* **13**, pp. 126–127.

L. LeCam (1973). Convergence of estimates under dimensionality restrictions, *Annals of Statistics* **1**, pp. 38–53.

E. J. G. Pitman (1979). *Some Basic Theory for Statistical Inference*, Chapman and Hall, London.

R. Rubinstein (1980). Generating random vectors uniformly distributed inside and on the surface of different regions, IBM Thomas J. Watson Research Center, Technical Report RC 8409.

R. Rubinstein (1981). *Simulation and the Monte Carlo Method*, Wiley, New York.

B. W. Schmeiser (1980). Random variate generation: a survey, Proceedings of the 1980 Winter Simulation Conference, Orlando, Florida.

B. W. Schmeiser and M. A. Shalaby (1980). Acceptance/rejection methods for beta variate generation, *Journal of the American Statistical Association* **75**, pp. 673–678.

B. W. Schmeiser and A. J. G. Babu (1980). Beta variate generation via exponential majorizing functions, *Operations Research* **28**, pp. 917–926.

R. J. Serfling (1979). A variation on Scheffé's theorem, with application to nonparametric density estimation, Report M502, Department of Statistics, Florida State University.

K. S. Shanmugam (1977). On a modified form of Parzen estimator for nonparametric pattern recognition, *Pattern Recognition* **9**, pp. 167–170.

M. Sibuya (1962). A method for generating uniformly distributed points on n-dimensional spheres, *Annals of the Institute of Statistical Mathematics* **44**, pp. 81–85.

Y. Tashiro (1977). On methods for generating uniform random points on the surface of a sphere, *Annals of the Institute of Statistical Mathematics* **29**, pp. 295–300.

J. R. Thompson and M. S. Taylor (1982). A data-based random number generator for a multivariate distribution, *Proceedings of the NASA Workshop on Density Estimation and Function Smoothing*, held at Texas A & M University, College Station, Texas, pp. 214–225.

J. von Neumann (1951). Various techniques in connection with random digits, *National Bureau of Standards AMS* **12**, pp. 36–38.

A. J. Walker (1977). An efficient method for generating discrete random variables with general distributions, *ACM Transactions on Mathematical Software* **3**, pp. 253–256.

CHAPTER 9

The Transformed Kernel Estimate

1. INTRODUCTION

The kernel estimate

$$f_n(x) = (nh)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{1}$$

has the disadvantage that h is not locally adjusted. This is reflected in the results of Chapter 5 where we have seen that the performance of the kernel estimate deteriorates when f becomes less smooth or heavy-tailed. To some extent, we can alleviate the problems by estimating the density of a transformed random variable, and then taking the inverse transform.

The *transformed kernel estimate* (Devroye et al., 1983) is based upon a transformation T: $R^1 \to [0,1]$ which is strictly monotonically increasing, continuously differentiable, one-to-one and onto, and which has a continuously differentiable inverse. The transformed data sequence is $Y_1, \ldots, Y_n$, where $Y_i = T(X_i)$. Note that Y_1 has density

$$g(x) = f(T^{-1}(x))T^{-1'}(x).$$

Now, g is estimated by g_n from $Y_1, \ldots, Y_n$, and f is estimated by

$$f_n(x) = g_n(T(x))T'(x). \tag{2}$$

The key observation is that if g_n is a density on $[0,1]$, the f_n is a density on R^1, and furthermore,

$$\int |f_n - f| = \int |g_n - g|.$$

In other words, the L_1 error is invariant under monotone tranformations. On the other hand, if g_n is a kernel estimate, the L_1 error is proportional to $B^*(g)$ as defined in Chapter 5. Thus, we should choose T so as to minimize $B^*(g)$. This leads to the natural question of the best form for g, which is answered in Theorem 5.3: $B^*(g)$ is always at least equal to $(2^9/3^4)^{1/5}$, and this minimum is attained for the isosceles triangular density on $[0,1]$. Because of this, we should choose T in such a way that g is isosceles triangular. If the distribution function F of f were known, this optimal transformation would be

$$T(x) = \begin{cases} \sqrt{F(x)/2}\,, & F(x) \le \frac{1}{2}, \\ 1 - \sqrt{(1-F(x))/2}\,, & F(x) > \frac{1}{2}. \end{cases} \tag{3}$$

From Table 5.1, we recall that the optimal choice for h is $(5/192\pi n)^{1/5}$ when g is triangular and K is the Epanechnikov kernel. Thus, g_n and f_n are completely defined if F is known. Unfortunately, F is not known and must be replaced by some estimate. Also, g_n is usually not a density on $[0,1]$ because some portions of g_n stick out beyond 1 or to the left of 0. To take care of this problem, we will use

$$\tilde{g}_n(x) = \frac{g_n(x)}{\int_0^1 g_n(y)\,dy} \tag{4}$$

instead of g_n. The integral does not cause computational problems because

$$\int_0^1 g_n(x) = \frac{1}{n}\sum_{i=1}^{n} \int_{(Y_i-1)/h}^{Y_i/h} K(x)\,dx. \tag{5}$$

If we define $f_n(x) = \tilde{g}_n(T(x))T'(x)$, then obviously, $\int |f_n - f| = \int |\tilde{g}_n - g|$. But our theoretical results on which we based our choice of g were valid for g_n. However, we are safe because for all g and g_n,

$$\int |\tilde{g}_n - g| \le \int |g_n - g| \tag{6}$$

(see Theorem 11.3).

The only unknown in the design at this moment is our transformation T. We point out that for a transformed histogram estimate, the optimal T gives a uniform $[0,1]$ density and should therefore be equal to $T(x) = F(x)$, all x. The h to be used in the histogram estimate is $(2\pi n)^{-1/3}$ (Table 5.1).

Because T is smooth, g inherits the smoothness properties of f. Thus, transformed estimates are probably not very good to take care of annoying discontinuities. They are mainly used for improvements in the performance of the original estimates due to better tail estimates. For example, if f is unimodal, then the optimal inverse transform will stretch the kernels in the tails. The visual effect is that of a method with a variable h depending upon x; h will usually "seem" larger in the tails, and smaller near the mode.

Another important factor is that $\tilde{g}_n$ is easy to plot because its support is compact. This point also led to the development of Parzen's density-quantile function estimate (Parzen, 1979).

Quite a few papers have been written about the choice of h, but, as we know, the expected L_1 error can decrease no faster than $n^{-2/5}$ times $B^*(f)$ times a constant. In this chapter, we suggest that rather than worry about h, we should try to work on $B^*(f)$, in a move to widen the horizons of the kernel estimate.

2. CHOOSING A TRANSFORMATION

Choosing a transformation is not a sinecure. In a vast number of applications, one suspects that f belongs to a certain family of densities (usually a parametric family), or at least is close to a given member of this family. If the family is a parametrized by θ, with distribution function F_θ, the natural approach is to estimate θ by $\tilde{\theta}$ in a *robust* manner, and use $F_{\tilde{\theta}}$ in the expression of the optimal transformation T. Throughout we use the same h, that is, the optimal h for the isosceles triangular density on $[0, 1]$.

Particularly attractive are the so-called "quick and dirty" robust estimates, based upon ideas given in Gastwirth (1966): for example, if $X_{(1)} < \cdots < X_{(n)}$ are the order statistics for $X_1, \ldots, X_n$, Gastwirth's estimate of the mean of a normal family is

$$\hat{\mu} = 0.3X_{(n/3)} + 0.4X_{(n/2)} + 0.3X_{(2n/3)}.$$

We refer to Huber (1972) and Andrews et al. (1972) for more examples of such simple robust estimates of location. For robust estimates of scale parameters, we could use the two-quantile method of Chapter 5, Section 6. For example, for the Cauchy family, this would give the following estimate of scale:

$$\hat{\sigma} = \tfrac{1}{2}\left(X_{(3n/4)} - X_{(n/4)}\right).$$

Except in trivial situations, the normal and Cauchy families are too small. For a survey of families with more parameters, see Schmeiser (1977). The

reason that we are reluctant to recommend nonparametric families is that the consistency and rate of convergence of the resulting estimate become harder to verify, and may even be questionable. With just a few parameters, T is nearly constant, and the consistency is not jeopardized.

3. ESTIMATION OF DENSITIES WITH LARGE TAILS

There are two factors that determine the efficiency of the kernel estimate: discontinuities or sharp oscillations, and large tails. The former factor, captured for smooth densities by $\int |f''|$, is infinite for densities with simple discontinuities such as the uniform density on $[0, 1]$. The latter factor, measured by $\int \sqrt{f}$, is infinite for densities with a large tail such as the Cauchy density. We have seen that when one or both of these factors is infinite, we must have $n^{2/5}E(J_n) \to \infty$ for the standard kernel estimate, regardless of the choice of h as a function of n.

The transformation to a triangular density eliminates the discontinuities and the tails, and should improve the performance of the estimate as measured by the L_1 error. In the inverse transformation, the discontinuities are reconstructed, creating the illusion that the transformed estimate works as a kernel estimate with locally adapted smoothing factor h. For example, h will usually seem bigger in the tails. This phenomenon will be illustrated with the aid of the notion of an isolated bump. It will partially explain the improvement that is obtained in the L_1 error.

An isolated bump in *any* density estimate is associated with one of the data points $X_1, \ldots, X_n$: X_i defines an isolated bump if there exists an interval $[a, b]$ with the property that $X_i \in [a, b]$, no other point X_j belongs to $[a, b]$, $\int_a^b f_n > 0$, and $f_n = 0$ on $[a - \varepsilon, a) \cup (b, b + \varepsilon]$ for some $\varepsilon > 0$. Assume, for example, that we are using the kernel estimate with Epanechnikov's kernel. Then X_i defines an isolated bump if and only if $[X_i - 2h, X_i + 2h]$ contains no data point except X_i. Thus, in the graph of f_n, $[X_i - h, X_i + h]$ appears as a separate hill, and it would seem that the data point "X_i" is wasted. Note also that the number of isolated bumps is invariant under strictly monotone transformations such as the ones considered in this chapter.

The total number of isolated bumps, B_n, bounds from below the number of hills in a graph. For example, when we are estimating a unimodal density, we would like the number of separate hills to be 1 and $B_n = 0$. As we will show in this section, this is usually not the case. For example, for the normal density with optimal h, $E(B_n)$ increases at least as $n^{1/5}/\sqrt{\log n}$, and the situation gets worse for longer-tailed densities. We will also show that for the triangular density, $E(B_n) = o(1)$.

The basic starting formula is

$$E(B_n) = nP([X_1 - 2h, X_1 + 2h] \text{ has no other data points})$$

$$= n\int f(x)\left(1 - \int_{x-2h}^{x+2h} f(y)\,dy\right)^{n-1} dx.$$

THEOREM 1 (General Result). *For all* f, $E(B_n) = o(n)$ *when* $h \to 0$, $nh \to \infty$.

Proof. We note that $E(B_n)/n = \int f(x) r_n(x)\,dx$, where the integrand $r_n(x)$ is $[0,1]$-valued and $r_n(x) \le \exp(-(n-1)\int_{x-2h}^{x+2h} f(y)\,dy) \to 0$ for almost all x (this follows from the fact that by the Lebesgue density theorem (Theorem 2.2) the exponent in the upper bound is asymptotic to $4nhf(x)$ for almost all x). Thus, Theorem 1 follows after an application of the Lebesgue dominated convergence theorem.

THEOREM 2 (Densities with a Regularly Varying Tail). *Let* f *be strictly monotonically decreasing on* $[0, \infty)$ *with uniquely defined inverse, and let* f *be* 0 *on* $(-\infty, 0)$ *for the sake of convenience. Assume further that* f *is regularly varying at* ∞ *with exponent* $r < -1$, *that is,*

$$\lim_{x \to \infty} \frac{f(tx)}{f(x)} = t^r, \quad \text{all } t > 0.$$

If $h \to 0$, $nh \to \infty$, *then*

$$E(B_n) \ge \frac{L(n)}{(nh)^{1/r}h}$$

for some slowly varying function L (*i.e., a regularly varying function with exponent* 0).

Proof. We will use the following facts:

(i) $\lim_{x \to \infty} xf(x)/\int_x^\infty f(y)\,dy = -r - 1$ (Dehaan, 1975, Theorem 1.2.1);

(ii) $f^{-1}(1/x)$ is $-1/r$ varying at ∞ (Dehaan, 1975, p. 22);

(iii) $f(x) = x^r L(x)$ for some slowly varying function L (Seneta, 1976, Lemma 2.1).

We will use the same symbol L for all our slowly varying functions. The set of all x for which $n\int_{x-2h}^{x+2h} f(y)\,dy \le \frac{1}{2}$ will be called A_n, and the set of all x for which $x > 2h$, $4nhf(x - 2h) < \frac{1}{2}$ will be called A_n^*. Then, Theorem 2

follows from

$$E(B_n) \geq n\int_0^\infty f(x)\left(1 - n\int_{x-2h}^{x+2h} f(y)\,dy\right)_+ dx$$

$$\geq \frac{n}{2}\int_{A_n} f \geq \frac{n}{2}\int_{A_n^*} f$$

$$\sim \frac{-n}{2(r+1)}\left(2h + f^{-1}\left(\frac{1}{8nh}\right)\right) f\left(2h + f^{-1}\left(\frac{1}{8nh}\right)\right)$$

$$\sim \frac{-n}{2(r+1)} f^{-1}\left(\frac{1}{8nh}\right) f\left(f^{-1}\left(\frac{1}{8nh}\right)\right)$$

$$= \frac{-n}{2(r+1)} f^{-1}\left(\frac{1}{8nh}\right)\frac{1}{8nh}$$

$$= \frac{L(8nh)(8nh)^{-1/r}}{-16(r+1)h}.$$

EXAMPLE. For the positive Student's t density $c/(1+x^2)^{(a+1)/2}$, $x \geq 0$, $a > 0$, we have $r = -(a+1)$, and thus

$$E(B_n) \geq L(n)h^{-1}(nh)^{1/(a+1)} = L(n)n^{1/(a+1)}h^{-a/(a+1)}.$$

For fixed $\varepsilon > 0$, $\beta \in (0,1)$, we can find an a such that $(1+a\beta)/(1+a) > 1 - \varepsilon/2$. Thus, for every $\varepsilon > 0$ and every sequence $h \sim c/n^\beta$, $\beta \in (0,1)$, there exists an a such that $E(B_n) \geq L(n)n^{1-\varepsilon/2} > n^{1-\varepsilon}$ for all n large enough. The last inequality follows from a property of slowly varying functions (see Seneta, 1976, p. 33). Thus, for any polynomially decreasing sequence h satisfying $h \to 0$, $nh \to \infty$, $E(B_n)$ can be forced to increase at any given polynomial rate $n^{1-\varepsilon}$ merely by choosing an appropriate density f in the Student's t family. In particular, when h decreases as $n^{-1/5}$, we have $E(B_n) \geq L(n)n^{(1+a/5)/(1+a)}$. The exponent in the last lower bound varies from 1 ($a \downarrow 0$) to $\frac{1}{5}$ ($a \to \infty$).

THEOREM 3 (The Isosceles Triangular Density). *For the isosceles triangular density on* $[0,1]$, *we have*

$$E(B_n) \leq (1 + o(1))\left((32nh^2)^{-1} + 32nh^2 e^{-8nh^2}\right)$$

when $h \to 0$, $nh \to \infty$. *In particular, when* $nh^2 \to \infty$, *we have* $E(B_n) \to 0$.

Proof. We start from

$$E(B_n) \le 2\int_0^{1/2} n \cdot 4x \exp\left(-(n-1)\int_{x-2h}^{x+2h} f(y)\,dy\right) dx.$$

We split the integral into three pieces, $[0, 2h]$, $[2h, \frac{1}{2} - 2h]$ and $[\frac{1}{2} - 2h, \frac{1}{2}]$. The contribution of the first piece does not exceed $8n \cdot 2h \exp(-(n-1) \cdot 2(2h)^2) \cdot 2h = (32 + o(1))nh^2 \exp(-8nh^2)$. The contribution of the third term is at most $2n \cdot 2 \cdot 2h \exp(-(n-1) \cdot 4h \cdot 4(\frac{1}{2} - 2h)) = (8 + o(1)) \cdot nh \exp(-8nh)$. The contribution of the second term is at most

$$2\int_{2h}^{\infty} 4nxe^{\ 16xh(n-1)}\,dx = \left(32h^2(n-1)\right)^{-1} \frac{n}{n-1} \int_{32(n-1)h^2}^{\infty} ye^{-y}\,dy$$

$$\le \frac{1 + o(1)}{32nh^2}.$$

This concludes the proof of Theorem 3.

By Theorem 3, the kernel estimate has with high probability no isolated bumps when f is triangular (in fact, $E(B_n) = O(n^{-3/5})$ for $h \sim n^{-1/5}$). The same is true for the transformed kernel estimate when the transformation is "perfect." Not only do we have a reduction in the number of isolated bumps, but also in the oscillation.

Let us finally note without proof that for the normal density, $E(B_n)$ is at least equal to a constant divided by $h\sqrt{\log(nh)}$ when $h \to 0$, $nh \to \infty$, and that thus with the optimal choice for h, $E(B_n)$ increases at least as $n^{1/5}/\sqrt{\log n}$. Thus, nothing would be gained by normalizing the data instead of triangularizing the data.

4. CONSISTENCY

For fixed transformations T: $R^1 \to R^1$ satisfying the conditions of Section 1, we have $J_n = \int |g_n - g|$, and thus certainly the exponential bound of Theorem 3.1 for J_n applies: $P(J_n > \varepsilon) \le \exp(-cn)$, all $n > n_0$, where $c > 0$ is a function of ε and n_0 is a number depending upon g and ε. Furthermore, the lower bound of Theorem 5.2 remains valid, but we can no longer be sure of the upper bound $C^*A(K)B^*(f)$ for $E(J_n)n^{2/5}$ (Theorem 5.1) as h may not be optimal for g. We do have a guarantee however that with the choice $h = (5/192\pi n)^{1/5}$, $E(J_n)$ decreases as $n^{-2/5}$ when $B^*(g) < \infty$. We also recall here the relative insensitivity of $E(J_n)$ to slightly suboptimal choices for h (see Section 5.6).

For variable transformations T, we must worry about the consistency of the resulting estimate.

The transformation $Y_i = T(X_i)$ is usually of the form

$$Y_i = T_n(X_i; X_1, \ldots, X_n),$$

for some Borel measurable function T_n satisfying the following conditions: T_n: $R^1 \to R^1$ is strictly monotonically increasing, one-to-one and onto, and continuously differentiable. Its inverse is also continuously differentiable.

Consider the transformed kernel estimate with Epanechnikov kernel K, and smoothing factor $h = \frac{1}{2}(5/6\sigma n)^{1/5}$ (which is optimal for the triangular density on $[0,1]$). We will not worry for the time being about transformations T_n: $R^1 \to [0,1]$ and the corresponding normalizations, because, as we have seen, this is an asymptotically negligible detail. We have the following densities:

f: density of $X_1, \ldots, X_n$ (the data).
g: density of $Y_i = T_n(X_i)$, given $X_1, \ldots, X_n$.
g^*: density of $T(X_1)$, where T is some given transformation.
g_n: transformed kernel estimate based upon $Y_1, \ldots, Y_n$.
g_n^*: transformed kernel estimate based upon $Z_i = T(X_i)$, $1 \le i \le n$.

We have the following inequality that can help us in proving the convergence to 0 of $\int |g_n - g|$:

$$\int |g_n - g| \le \int |g^* - g| + \int |g_n^* - g^*| + 6 \sup_x |T_n(x) - T(x)|.$$

To see this, note that K is Epanechnikov's kernel, and thus that

$$\int |g_n - g_n^*| \le (nh)^{-1} \int \sum_{i=1}^{n} \left| K\left(\frac{y - Y_i}{h}\right) - K\left(\frac{y - Z_i}{h}\right) \right| dy$$

$$\le \frac{1}{n} \sum_{i=1}^{n} \frac{3}{2} \cdot 4|Y_i - Z_i| \le 6 \sup_x |T_n(x) - T(x)|.$$

The transformation T is in some sense the limit of T_n. For example, when T_n is obtained by estimating some parameters, the actual form of T is known. One would hope that of the three terms on the right-hand side of the inequality, the middle term dominates (it usually will), so that we can in effect replace T_n for rate of convergence studies by the fixed transformation T.

One of the terms giving some trouble is $\int |g^* - g|$. We know that it tends to 0 in probability when it does so at almost every x (Theorem 2.8). But this is the case when f is a.e. continuous, $T_n^{-1} \to T^{-1}$ a.e. and $T_n^{-1'} \to T^{-1'}$ a.e. in probability.

REFERENCES

D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey (1972). *Robust Estimates of Location: Survey and Advances*, Princeton University Press, Princeton.

L. Dehaan (1975). *On Regular Variation and Its Applications to the Weak Convergence of Sample Extremes*, Mathematisch Centrum Tracts 32, Mathematisch Centrum, Amsterdam.

L. Devroye, F. Machell, and C. S. Penrod (1983). The transformed kernel estimate, Technical Report, Applied Research Laboratories, University of Texas, Austin, Texas.

J. L. Gastwirth (1966). On robust procedures, *Journal of the American Statistical Association* **61**, pp. 929–948.

P. J. Huber (1972). Robust statistics: a review, *Annals of Mathematical Statistics* **43**, pp. 1041–1067.

E. Parzen (1979). Nonparametric statistical data modeling, *Journal of the American Statistical Association* **74**, pp. 105–131.

B. W. Schmeiser (1977). Methods for modelling and generating probabilistic components in digital computer simulation when the standard distributions are not adequate: a survey, *Proceedings of the Winter Simulation Conference*, pp. 51–55.

E. Seneta (1976). *Regularly Varying Functions*, Lecture Notes in Mathematics 508, Springer-Verlag, Berlin.

CHAPTER 10

Applications in Discrimination

1. THE DISCRIMINATION PROBLEM

The problem of discrimination (pattern classification, statistical pattern recognition) is usually formulated as follows: the observation X is a random variable taking values in R^d, and the label Y is a random variable taking values in $\{1, \ldots, M\}$. Given X, one has to guess the value of Y, and this is called a decision. The decision is a measurable function: g: $R^d \to \{1, \ldots, M\}$, and the probability of error is $P(g(X) \neq Y)$. If the distribution of (X, Y) can be characterized by the probability measure for X, μ, and the regression functions (also called a posteriori probabilities,

$$p_i(x) = P(Y = i | X = x), \qquad x \in R^d, 1 \leq i \leq M.$$

A decision g^* is called Bayesian if

$$p_{g^*(x)}(x) = \max_i p_i(x) \quad \text{almost all } x(\mu). \tag{1}$$

If X has a density f and has conditional densities f_i given $Y = i$, $1 \leq i \leq M$, then

$$p_i(x) = \frac{p_i f_i(x)}{f(x)}, \quad \text{almost all } x(\mu),$$

where $p_i = P(Y = i)$. Thus, for a Bayesian decision we have

$$p_{g^*(x)} f_{g^*(x)}(x) = \max_i p_i f_i(x), \quad \text{almost all } x(f). \tag{2}$$

In discrimination the problem is to minimize the probability of error if the p_i's and f_i's are unknown, and a sample $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$

of independent identically distributed copies of (X, Y) is available. We assume that D_n and (X, Y) are independent. Y is now estimated by $g_n(X)$, a measurable function of X and D_n (the dependence upon D_n is suppressed in the notation), and the quantity of interest is the conditional probability of error

$$L_n = P(g_n(X) \neq Y | D_n). \tag{3}$$

In particular, we would like to find sequences of functions g_n for which

$$L_n \to L^* = \min_g P(g(X) \neq Y) \quad \text{almost surely.} \tag{4}$$

Here L^* is usually called the Bayes probability of error. This is precisely what we will describe in this chapter, several ways of choosing such sequences. We will not take a deep look at the properties of these sequences beyond (4). Rather, we would like to point out how the results of Chapters 1–9 on density estimation can be applied to obtain (4). In particular, we will assume that X has a density f. It is stressed however that most of the results stated in this chapter remain valid for all probability measures μ on the Borel sets of R^d.

To approximate the Bayesian decision, we could first estimate all p_i's by [0, 1]-valued functions of D_n, say $\tilde{p}_i$, $1 \leq i \leq M$, and let g_n satisfy

$$\tilde{p}_{g_n(x)}(x) = \max_i \tilde{p}_i(x). \tag{5}$$

If X has a density f, and $p_i f_i(x)$ is estimated from D_n by $\tilde{p}_i \tilde{f}_i(x)$, then g_n may be defined by

$$\tilde{p}_{g_n(x)} \tilde{f}_{g_n(x)}(x) = \max_i \tilde{p}_i \tilde{f}_i(x). \tag{6}$$

THEOREM 1.

(i) *If g^* is a Bayesian decision, then*

$$L^* = P(g^*(X) \neq Y).$$

(ii) *If g_n satisfies* (5), *then*

$$0 \leq L_n - L^* \leq \sum_{i=1}^{M} \int |p_i(x) - \tilde{p}_i(x)| \mu(dx).$$

(iii) *If X has a density and g_n satisfies* (6), *then*

$$0 \leq L_n - L^* \leq \sum_{i=1}^{M} \int |p_i f_i(x) - \tilde{p}_i \tilde{f}_i(x)| \, dx.$$

REMARK 1. Various versions of relations (ii) and (iii) were proved by Van Ryzin (1966), Wolverton and Wagner (1969), Csibi (1975), Györfi (1974, 1978), Devroye and Wagner (1976), and Devroye (1982b).

Proof. By (1),

$$P(g^*(X) \neq Y) = 1 - \sum_{i=1}^{M} P(Y = i, g^*(X) = i)$$

$$= 1 - \sum_{i=1}^{M} \int_{[g^*(x)=i]} p_i(x)\mu(dx)$$

$$= 1 - \int \max_i p_i(x)\mu(dx) \leq 1 - \int p_{g(x)}(x)\mu(dx) \quad (7)$$

for any $g: R^d \to \{1, \ldots, M\}$. Take the infimum over all g, and (i) follows. Also,

$$L_n = 1 - \sum_{i=1}^{M} \int_{[g_n(x)=i]} p_i(x)\mu(dx) = 1 - \int p_{g_n}(x)\mu(dx), \quad (8)$$

so that by combining (7), (i), and (8) we have

$$L_n - L^* = \int \Big(\max_i p_i(x) - p_{g_n(x)}(x) \Big) \mu(dx)$$

$$= \int \Big(\max_i p_i(x) - \max_i \tilde{p}_i(x) \Big) \mu(dx)$$

$$+ \int \Big(\tilde{p}_{g_n(x)}(x) - p_{g_n(x)}(x) \Big) \mu(dx)$$

$$\leq \sum_{i=1}^{M} \int |p_i(x) - \tilde{p}_i(x)| \mu(dx), \quad (9)$$

and (ii) is proved. Now, (iii) is a simple consequence of (ii) for $\tilde{p}_i(x) = \tilde{p}_i \tilde{f}_i(x)/f(x)$.

2. SLOW RATES OF CONVERGENCE

It is quite interesting that the L_1 errors of the density estimates $\tilde{f}_i$ provide an upper bound for the probability of error. But there is also a converse

relationship of sorts: for example, the probability of error (L_n) can tend to L^* at an arbitrary slow rate (Devroye, 1982a).

THEOREM 2. *Let a_n be a sequence of positive numbers tending to* 0, *let* $M = 2$, *and* $c \in [0, \frac{1}{2})$. *Let g_n be arbitrary. Then there exists a distribution of* (X, Y) *for which* X *is uniformly distributed on* $[0,1]$ *and* $L^* = c$, *so that*

$$\limsup_{n \to \infty} \frac{E(L_n) - c}{a_n} = \infty.$$

Proof. The proof follows the general lines of the proof of (ii) of Theorem 4.1. We will merely outline the construction of the randomized family for the special case $c = 0$. Let X have a uniform density on $[0,1]$, and let $b = 0 \cdot b_1 b_2 b_3 \cdots \in [0,1]$ (the b_i's are the coefficients in a binary expansion of b). Then define

$$p_2(x) = f_b(x) = \sum_{i=1}^{\infty} b_i I_{[x_i, x_{i+1})}(x), \qquad p_1(x) = 1 - p_2(x),$$

$$Y = 1 + f_b(X), \qquad Y_n = 1 + f_b(X_n),$$

where $0 = x_1 \le x_2 \le \cdots \le x_n \uparrow 1$. We define $q_i = x_{i+1} - x_i$, $i = 1, 2, \ldots$. Assume that $B = 0 \cdot B_1 B_2 \cdots$ is uniformly distributed on $[0,1]$ and independent of $X, X_1, X_2, \ldots, X_n$, and let us use the notation $R_n(b) = E(L_n)$, $b \in [0,1]$. We have

$$\begin{aligned}
\sup_{b \in [0,1]} R_n(b) &\ge E\big(R_n(B)\big) = P\big(g_n(X, D_n) \ne Y\big) \\
&= E\big(P\big(g_n(X, D_n) \ne Y | X, X_1, \ldots, X_n\big)\big) \\
&\ge E\big(I_A P\big(g_n(X, D_n) \ne Y | X, X_1, \ldots, X_n\big)\big),
\end{aligned}$$

where A is the event $\bigcap_{i=1}^n [f_B(X) f_B(X_i) = 0]$. On A, Y and $g_n(X, X_1, Y_1, \ldots, X_n, Y_n)$ are independent given $X, X_1, \ldots, X_n$. Thus, on A, $P(g_n(X, D_n) \ne Y | X, X_1, \ldots, X_n) = \frac{1}{2}$, and therefore

$$\sup_{b \in [0,1]} R_n(b) \ge \frac{1}{2} P\left(\bigcap_{i=1}^n \big[f_B(X) f_B(X_i) = 0 \big] \right) = \frac{1}{2} \sum_{i=1}^{\infty} q_i (1 - q_i)^n.$$

The rest of the proof is the same as that of (ii) of Theorem 4.1.

Theorem 2 implies that rate of convergence results for $E(L_n)$ can only exist under some smoothness assumptions about the regression function.

Rather than pursuing the issue of the best possible rate of convergence for certain families of distributions of (X, Y), we will concentrate on consistency results for the most popular nonparametric discrimination methods.

3. THE KERNEL METHOD IN DISCRIMINATION

Let K be a function in $L_1(R^d)$ with $\int K = 1$, and consider the modified kernel density estimate

$$\tilde{p}_i \tilde{f}_i(x) = \left(nh_n^d\right)^{-1} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h_n}\right) I_{[Y_j = i]}, \qquad 1 \le i \le M. \tag{10}$$

The corresponding decision is defined in (6). Under some additional conditions on K, Devroye and Wagner (1980) and Spiegelman and Sacks (1980) showed that $L_n \to L^*$ in probability as $n \to \infty$ when $h_n \to 0$, $nh_n^d \to \infty$. In Devroye (1981), strong convergence was obtained under the additional condition $nh_n^d / \log n \to \infty$. In all these papers, there are no conditions on the distribution of (X, Y). Theorem 3 below is valid whenever X has a density: it states that L_n converges to L^* exponentially. The conditions of convergence cannot be improved upon.

THEOREM 3. *If X has a density, $h_n \to 0$ and $nh_n^d \to \infty$, then the kernel method defined by* (6) *and* (10) *satisfies*:

For all $\varepsilon \in (0,1)$ there exists $n_0 > 0$ such that
$$P(L_n - L^* > \varepsilon) \le \exp\left(-c_1 n\varepsilon^2\right), n \ge n_0.$$

Here $c_1 > 0$ is a constant depending upon K only.

Proof. Because of Theorem 1, one only has to show that

$$P\left(\sum_{i=1}^{M} \int |\tilde{p}_i \tilde{f}_i(x) - p_i f_i(x)| dx > \varepsilon\right) \le \exp\left(-c_1 n\varepsilon^2\right), \qquad n \ge n_0.$$

Because $E(\tilde{p}_i \tilde{f}_i(x)) = p_i K_{h_n} * f_i(x)$, we have by Theorem 2.1, $\int |p_i f_i - E(\tilde{p}_i \tilde{f}_i)| \to 0$ for all i. Thus, it is sufficient to establish the exponential inequality for

$$P\left(\sum_{i=1}^{M} \int |\tilde{p}_i \tilde{f}_i(x) - E(\tilde{p}_i \tilde{f}_i(x))| dx > \varepsilon\right).$$

As in the proof of Lemma 3.2, we need only consider kernels K that are indicators of rectangles A. Introduce the measures μ_i and μ_{ni} defined on Borel sets B as follows:

$$\mu_i(B) = P(Y_1 = i, X_1 \in B); \quad \mu_{ni}(B) = \frac{1}{n} \sum_{j=1}^{n} I_{[Y_j = i,\, X_j \in B]},$$

$$1 \le i \le M.$$

Then

$$\sum_{i=1}^{M} \int |\tilde{p}_i \tilde{f}_i(x) - E(\tilde{p}_i \tilde{f}_i(x))|\, dx$$

$$= \sum_{i=1}^{M} \int |\mu_{ni}(x + h_n A) - \mu_i(x + h_n A)|\, dx.$$

This can be treated by technique of Lemma 3.2 and the rest of the proof is straightforward. We conjecture that Theorem 3 remains valid for *all* distributions of (X, Y)

4. HISTOGRAM-BASED DISCRIMINATION

Next, we consider histogram-based decisions in which R^d is partitioned into sets $A_{n1}, A_{n2}, \ldots$, and the estimates in (6) are of the form

$$\tilde{p}_i \tilde{f}_i(x) = \frac{1}{n} \sum_{m=1}^{n} \frac{I_{[X_m \in A_{nj},\, Y_m = i]}}{\lambda(A_{nj})}, \qquad x \in A_{nj}. \tag{11}$$

THEOREM 4. *If X has a density, the sequence of partitions satisfies* (3.13)–(3.15), *and decision* (6) *is used with the histogram estimate* (11), *then the following statement is valid*:

For each $\epsilon \in (0,1)$ there exists $n_0 > 0$ such that
$P(L_n - L^* > \epsilon) \le \exp(-c_2 n \epsilon^2), n \ge n_0.$

Here $c_2 > 0$ is a universal constant.

The proof is a copy of the proof of Lemma 3.4 and will not be given here. The complete convergence of L_n to L^* for all distributions of (X, Y) was

obtained by Devroye and Györfi (1983). The conditions of convergence are essentially the same as those of weak convergence indicating once again the equivalence of all types of convergence. For histogram-based decisions in which the partition of R^d depends upon the data, and corresponding weak convergence results, see Gordon and Olshen (1978) and the references found there.

5. THE NEAREST-NEIGHBOR METHOD

Another popular nonparametric decision is based upon the notion of k nearest neighbors (Cover and Hart, 1967). Given X, D_n is permuted according to increasing values of $\|X_i - X\|$: a vector of ranks is obtained $(R_1(X), \ldots, R_n(X))$, where $X_{R_i(X)}$ is the ith nearest neighbor of X. Ties are broken by comparing indices. We note here that if X has a density f, ties occur with probability zero. The decision is based upon a majority vote among $Y_{R_j(X)}, 1 \le j \le k_n$, where k_n is a sequence of integers. Stone (1977) also considers the case of weighted voting: the ith nearest neighbor carries a weighted vote v_{ni}; for each class the total vote is computed, and the winning class is our decision. For the decision with equal weights among the k_n nearest neighbors, we note that it is equivalent to using the nearest-neighbor density estimate (Fix and Hodges, 1951, 1952; Loftsgaarden and Quesenberry, 1965) in (6):

$$\tilde{p}_i \tilde{f}_i(x) = \frac{1}{n} \sum_{j=1}^{k_n} \frac{I_{[Y_{R_j}(x)=i]}}{\lambda\left(S_{x,\|x-X_{R_{k_n}}(x)\|}\right)}. \tag{12}$$

This is due to the fact that the denominator in (12) is the same for all i. We will prove the following theorem:

THEOREM 5. *If X has a density f, $k_n \to \infty$, and $k_n/n \to 0$, then the following is true for the decision* (6) *based on the nearest-neighbor estimate* (12):

> *For each $\varepsilon \in (0,1)$ there exists $n_0 > 0$ such that*
> $P(L_n - L^* > \varepsilon) \le \exp(-c_3 n\varepsilon^2)$, $n \ge n_0$, *where $c_3 > 0$ depends upon the dimension only.*

Before we present the proof, a brief historical remark is in order. Stone (1977) showed that $L_n \to L^*$ in probability for all distributions of (X, Y) when $k_n \to \infty$ and $k_n/n \to 0$. We note here that these conditions on k_n are necessary. Proceeding via pointwise convergence, Devroye (1982b) showed that $L_n \to L^*$ almost surely for all distributions of (X, Y) when $k_n/n \to 0$ and $k_n/\log\log n \to \infty$. Beck (1979) proved Theorem 5 under some additional smoothness conditions on f and the f_i's. The proof given here is shorter and more general. It also imposes no conditions on f. Thus, the following statements are equivalent:

(i) $\lim_{n\to\infty} k_n = \infty$, $\lim_{n\to\infty} k_n/n = 0$.

(ii) $L_n \to L^*$ in probability whenever X has a density.

(iii) $L_n \to L^*$ exponentially whenever X has a density, that is, for each $\varepsilon > 0$ there exists $c > 0$ such that $P(L_n - L^* > \varepsilon) \le e^{-cn}$ for all n.

We would also like to point out that there is no hope of extending the exponential inequality of Theorem 5 to all distributions of (X, Y) (while keeping the conditions on k_n). For example, let X be 0 with probability 1, and let Y be 1 or 2 with probabilities $\frac{1}{3}$ and $\frac{2}{3}$, respectively. Because $(Y_{R_1(x)}, \ldots, Y_{R_{k_n}(x)})$ is $(Y_1, \ldots, Y_{k_n})$, we have

$$L_n = \tfrac{1}{3} + \tfrac{1}{3} I_A, \qquad L^* = \tfrac{1}{3},$$

where A is the event $[(1/k_n)\sum_{i=1}^{k_n} I_{[Y_i=1]} > \frac{1}{2}]$. Thus, by Kolmogorov's exponential lower bound (Stout, 1974, p. 262) (or Lemma 6.6), we have $P(L_n - L^* \ge \varepsilon) \ge \exp(-ck_n)$ for all n and some $c > 0$.

The proof of Theorem 5 requires a geometric property also applied by Fritz (1975) and Stone (1977). A *cone* of angle θ centered at x is defined as the collection of all points $y \in R^d$ such that angle $(y - x, z - x) < \theta$ for a given point $z \in R^d$. Thus, x, z, and θ determine the cone. Now, choose θ so small that for each $v \in \mathrm{Cone}(x, z, \theta)$,

$$\mathrm{Cone}(x, z, \theta) \cap S_{x, \|x-v\|} \subseteq S_{v, \|x-v\|}. \tag{13}$$

After having fixed θ, we define the integer M_d as the minimal number of cones of the form $\mathrm{Cone}(x, z_i, \theta), 1 \le i \le M_d$, needed to cover R^d.

LEMMA 1. *Let μ be a probability measure on R^d, and let*

$$B_a(x) = \left\{ z : \mu\left(S_{z, \|x-z\|}\right) \le a \right\}, \qquad x \in R^d.$$

Then

$$\mu\big(B_a(x)\big) \le M_d a.$$

Proof. Let $C_i, 1 \le i \le M_d$, be a collection of cones of the form $\text{Cone}(x, z_i, \theta)$ covering R^d and satisfying property (13). Now,

$$\mu\big(B_a(x)\big) \le \sum_{i=1}^{M_d} \mu\big(C_i \cap B_a(x)\big). \tag{14}$$

For fixed i, choose an arbitrary point $y \in C_i \cap B_a(x)$. By (13),

$$\mu\big(S_{x,\|x-y\|} \cap C_i \cap B_a(x)\big) \le \mu\big(S_{y,\|y-x\|}\big) \le a, \tag{15}$$

where we used the fact that $y \in B_a(x)$. Since y was arbitrary, we have

$$\mu\big(C_i \cap B_a(x)\big) \le a. \tag{16}$$

The lemma follows from (14) and (16).

Proof of Theorem 5. We introduce the notation

$$\tilde{p}_i(x) = \frac{1}{k} \sum_{j=1}^{k} I_{[Y_{R_j}(x)=i]}$$

and

$$p_i^*(x) = \frac{1}{k} \sum_{j=1}^{n} I_{[Y_j=i]} I_{[\|X_j-x\| \le r_n(x)]},$$

where $k = k_n$, and $r_n(x)$ is a solution of the equation

$$\frac{k}{n} = \mu\big(S_{x,r_n(x)}\big). \tag{17}$$

Note that decision (5) with $\tilde{p}_i(x)$ as defined above is equivalent to decision (6) with (12). We note that the solution $r_n(x)$ is positive since μ has a density. Also, $k/n \to 0$ implies that $r_n(x) \to 0$ for almost all $x(\mu)$. If C_d is the Lebesgue measure of the unit sphere of R^d, then (17) is equivalent to

$$k = n r_n^d(x) C_d \frac{\mu\big(S_{x,r_n(x)}\big)}{\lambda\big(S_{x,r_n(x)}\big)}. \tag{18}$$

Thus, by Theorem 2.2, and $k \to \infty$, we have $nr_n^d(x) \to \infty$ for almost all $x(\mu)$.

Obviously,

$$|p_i(x) - \tilde{p}_i(x)| \le |p_i(x) - E(p_i^*(x))| + |E(p_i^*(x)) - p_i^*(x)| + |p_i^*(x) - \tilde{p}_i(x)|. \quad (19)$$

We have

$$E(p_i^*(x)) = \frac{P(\|X_1 - x\| \le r_n(x), Y_1 = i)}{k/n} = \int_{S_{x,r_n(x)}} \frac{p_i(z)\mu(dz)}{\mu\left(S_{x,r_n(x)}\right)} \to p_i(x) \quad \text{for almost all } x(\mu), \quad (20)$$

by a slight generalization of Theorem 2.2. Thus, by the Lebesgue dominated convergence theorem, $\int |E(p_i^*(x)) - p_i(x)|\mu(dx) \to 0$ for all i.

Next, let μ_n be the empirical measure for $X_1, \ldots, X_n$, and let i be an integer in $\{1, \ldots, M\}$. Then,

$$\int |p_i^*(x) - E(p_i^*(x))|\mu(dx) = \int \frac{|\nu_n\left(S_{x,r_n(x)}\right) - \nu\left(S_{x,r_n(x)}\right)|}{\mu\left(S_{x,r_n(x)}\right)} \mu(dx),$$

where

$$\nu_n(A) = \frac{1}{n} \sum_{j=1}^{n} I_{[X_j \in A, Y_j = i]}$$

and

$$\nu(A) = E(\nu_n(A)), \quad A \text{ Borel set of } R^d.$$

Let h be $(k/n)^{1/d}$, and let $\mathscr{P}$ be a cubic partition of R^d with cubes of size h/N for some large integer N to be determined. The members of $\mathscr{P}$ are denoted by B, and the centers of these sets B are called b. T is a fixed

sphere $S_{0,r}$ for some large r. We have

$$\int |p_i^*(x) - E(p_i^*(x))|\mu(dx)$$

$$\leq \int \sum_{\substack{B: B\cap T\neq\varnothing \\ B\subseteq S_{x,r_n(x)}}} \left(\frac{|\nu_n(B)-\nu(B)|}{\mu\left(S_{x,r_n(x)}\right)}\right)\mu(dx)$$

$$+ \int \sum_{\substack{B: B\cap T=\varnothing \\ B\subseteq S_{x,r_n(x)}}} \left(\frac{\nu_n(B)+\nu(B)}{\mu\left(S_{x,r_n(x)}\right)}\right)\mu(dx) \tag{21}$$

$$+ \int \sum_{\substack{B: B\cap S_{x,r_n(x)}\neq\varnothing \\ B\cap S^c_{x,r_n(x)}\neq\varnothing}} \frac{\left(\nu_n\left(B\cap S_{x,r_n(x)}\right)+\nu\left(B\cap S_{x,r_n(x)}\right)\right)}{\mu\left(S_{x,r_n(x)}\right)}\mu(dx).$$

Applying Lemma 1 with $a = k/n$, the first term on the right-hand side of (21) is at most

$$\sum_{B: B\cap T\neq\varnothing} |\nu_n(B)-\nu(B)| \int \left(\frac{I_{S_{x,r_n(x)}}(b)}{\mu\left(S_{x,r_n(x)}\right)}\right)\mu(dx)$$

$$\leq M_d \sum_{B: B\cap T\neq\varnothing} |\nu_n(B)-\nu(B)|. \tag{22}$$

This tends to 0 exponentially, but the exponent depends upon M_d (see Lemma 3.1). The second term on the right-hand side of (21) is not greater than

$$M_d(\mu_n(T^c)+\mu(T^c)). \tag{23}$$

For a given $\varepsilon > 0$, we can first choose r large enough so that $\mu(T^c) < \varepsilon/M_d$. Then $P((23) > 3\varepsilon) \leq P(|\mu_n(T^c)| > \varepsilon/M_d) \leq 2\exp(-2n(\varepsilon/M_d)^2)$ by Hoeffding's inequality (Hoeffding, 1963). Thus, we need only look at the last term of (21).

To bound the last term of (21) from above, we will use the notation $A = S_{x,r_n(x)}$, $A^* = S_{x,(r_n(x)-h/N)_+}$, and $A_N = \{x:\ (C_d f(x))^{1/d} > \sqrt{N}\}$,

where $(\ )_+$ is the function that takes the positive part. We bound the term by

$$\int\left(\frac{\mu_n(A)-\mu_n(A^*)}{\mu(A)}\right)\mu(dx)+\int\left(\frac{\mu(A)-\mu(A^*)}{\mu(A)}\right)\mu(dx)$$

$$=\int\left(\int\left(\frac{I_A(z)-I_{A^*}(z)}{\mu(A)}\right)\mu(dx)\right)\mu_n(dz)$$

$$+\int\left(\frac{\mu(A)-\mu(A^*)}{\mu(A)}\right)\mu(dx). \tag{24}$$

The first term on the right-hand side of (24) can be written as $\int\zeta(z)\mu_n(dz)$, where $|\zeta(z)|\le M_d$. The last term is then $\int\zeta(z)\mu(dz)$. By Hoeffding's inequality, we see that $|\int\zeta(z)\mu_n(dz)-\zeta(z)\mu(dz)|\to 0$ exponentially with exponent depending upon M_d. Thus, we need only show that $\int\zeta(z)\mu(dz)$ (the last term of (24)) can be made arbitrarily small.

Choose N such that $\mu(A_N)<\varepsilon$, $d/\sqrt{N}<\varepsilon$, where $\varepsilon>0$ is an arbitrary number. In view of (18) and the definition of h, we have

$$(h/r_n(x))^d\to C_d f(x),\quad \text{almost all } x. \tag{25}$$

Thus, for almost all $x\notin A_N$, $(1-h/Nr_n(x))_+\to(1-(C_df(x))^{1/d}/N)_+\ge(1-1/\sqrt{N})_+>0$, and $r_n(x)>h/N$ for all n large enough. For such x,

$$\frac{\mu(A)-\mu(A^*)}{\mu(A)}\to 1-\left(1-\frac{(C_df(x))^{1/d}}{N}\right)^d\le\frac{d}{\sqrt{N}}<\varepsilon. \tag{26}$$

From this and the Lebesgue dominated convergence theorem, we deduce that

$$\limsup_{n\to\infty}\int\left(\frac{\mu(A)-\mu(A^*)}{\mu(A)}\right)\mu(dx)$$

$$\le\mu(A_N)+\limsup_{n\to\infty}\int_{A_N^c}\left(\frac{\mu(A)-\mu(A^*)}{\mu(A)}\right)\mu(dx)<2\varepsilon. \tag{27}$$

Thus, $\int|p_i^*(x)-E(p_i^*(x)|\mu(dx)$ tends to 0 exponentially.

We will finally consider the last term in (19). Again, for fixed $i \in \{1, \ldots, M\}$, we have almost surely

$$
\begin{aligned}
&|p_i^*(x) - \tilde{p}_i(x)| \\
&\quad = \frac{1}{k}\left| \sum_{j=1}^{n} I_{[Y_j = i, \|X_j - x\| \le r_n(x)]} - \sum_{j=1}^{n} I_{[Y_j = i, \|X_j - x\| \le \|X_{R_k(x)} - x\|]} \right| \\
&\quad \le \frac{1}{k} \sum_{j=1}^{n} |I_{[\|X_j - x\| \le r_n(x)]} - I_{[\|X_j - x\| \le \|X_{R_k(x)} - x\|]}| \\
&\quad = \left| \frac{1}{k} \sum_{j=1}^{n} I_{[\|X_j - x\| \le r_n(x)]} - 1 \right|. \qquad (28)
\end{aligned}
$$

If we consider a new discrimination problem with data (X_j, W_j), where $W_j = i$ for all j, and if we let the corresponding $p_i^*(x)$ be $q_i^*(x)$, then (28) implies that

$$
\int |p_i^*(x) - \tilde{p}_i(x)| \mu(dx) \le \int |q_i^*(x) - E(q_i^*(x))| \mu(dx), \qquad (29)
$$

which we know converges exponentially to 0.

To verify that the exponent of convergence does not depend upon M, proceed as follows: sum over all i on the left-hand side of (21). This gives, in the first instance, an expression as in (22) with a summation over all i, to which Lemma 3.1 can be applied. To handle the second term on the right-hand side of (21), absolutely no modifications are necessary, and similarly for the last term on the right-hand side of (21). For the last term on the right-hand side of (19), we begin with adding a summation sign to the first two lines of (28). The inequality in (28) remains valid as stated. Finally, add the summation over all i to both sides of (29). This completes the proof.

REFERENCES

J. Beck (1979). The exponential rate of convergence of error for k_n – NN nonparametric regression and decision, *Problems of Control and Information Theory* **8**, pp. 303–312.

T. M. Cover and P. E. Hart (1967). Nearest neighbor pattern classification, *IEEE Transactions on Information Theory* **IT-13**, pp. 21–27.

S. Csibi (1975). *Stochastic Processes With Learning Properties*, Springer-Verlag, Berlin.

L. Devroye (1981). On the almost everywhere convergence of nonparametric regression function estimates, *Annals of Statistics* **9**, pp. 1310–1319.

L. Devroye (1982a). Any discrimination rule can have an arbitrarily bad probability of error for finite sample size, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-4**, pp. 154–157.

L. Devroye (1982b). Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **61**, pp. 467–481.

L. Devroye and L. Györfi (1983). Distribution-free exponential bound on the L_1 error of partitioning estimates of a regression function, in *Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics*, G. Pflug, W. Grossmann, and W. Wertz (Eds.), D. Reidel, Hingham, MA.

L. Devroye and T. J. Wagner (1976). Nonparametric discrimination and density estimation, Technical Report 183, Electronics Research Centre, University of Texas, Austin, Texas.

L. Devroye and T. J. Wagner (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation, *Annals of Statistics* **8**, pp. 231–239.

E. Fix and J. L. Hodges (1951). Discriminatory analysis, nonparametric discrimination, consistency properties, Report No. 4, Project 21-49-004, School of Aviation Medicine, Randolph Field, Texas.

E. Fix and J. L. Hodges (1952). Nonparametric discrimination: small sample performance, Report No. 11, Project 21-49-004, School of Aviation Medicine, Randolph Field, Texas.

J. Fritz (1975). Distribution-free exponential error bound for nearest neighbor pattern classification, *IEEE Transactions on Information Theory* **IT-21**, pp. 552–557.

L. Gordon and R. A. Olshen (1978). Asymptotically efficient solutions to the classification problem, *Annals of Statistics* **6**, pp. 515–533.

L. Györfi (1974). Estimation of probability density and optimal decision function in RKHS, in *Progress in Statistics*, J. Gani, K. Sarkadi, and I. Vincze (Eds.), North-Holland, Amsterdam, pp. 281–301.

L. Györfi (1978). On the rate of convergence of nearest neighbor rules, *IEEE Transactions on Information Theory* **IT-24**, pp. 509–512.

W. Hoeffding (1963). Probability inequalities for sums of bounded random variables, *Journal of the American Statistical Association* **58**, pp. 13–30.

D. O. Loftsgaarden and C. P. Quesenberry (1965). A nonparametric estimate of a multivariate density function, *Annals of Mathematical Statistics* **36**, pp. 1049–1051.

C. Spiegelman and J. Sacks (1980). Consistent window estimation in nonparametric regression, *Annals of Statistics* **8**, pp. 240–246.

C. J. Stone (1977). Consistent nonparametric regression, *Annals of Statistics* **5**, pp. 595–645.

W. F. Stout (1974). *Almost Sure Convergence*, Academic Press, New York.

J. Van Ryzin (1966). Bayes risk consistency of classification procedures using density estimation, *Sankhya* **28**, pp. 261–270.

C. T. Wolverton and T. J. Wagner (1969). Asymptotically optimal discriminant functions for pattern classifications, *IEEE Transactions on Information Theory* **IT-15**, pp. 258–265.

CHAPTER 11

Operations on Density Estimates

The central theme of this chapter is the connection between the L_1 error of some density estimate and the L_1 error of the same density estimate after it has passed through operations, for example, taking the marginal density, forming product densities, convolving densities, truncating densities, and forming the nonnegative projection of a density are all rather common operations. We will wherever possible establish useful inequalities.

1. MARGINAL DENSITIES

Let f^* and g^* be densities on R^d (we are thinking here in general of an unknown density f^* and a sample-based estimate g^*, but the randomness implicit in g^* will be unimportant). Let f and g be the marginal densities on a subspace R^s of R^d. Then, we have the following theorem:

THEOREM 1.

$$\int_{R^s} |f - g| \leq \int_{R^d} |f^* - g^*|.$$

Proof. With a little abuse of notation, we have

$$\int_{R^s} |f - g| = \int_{R^s} \left| \int_{R^{d-s}} f^* - \int_{R^{d-s}} g^* \right| \leq \int_{R^d} |f^* - g^*|.$$

Theorem 1 is often nearly vacuous, because there exist examples with $d = 2$, $s = 1$, for which $f = g$ but $\int |f^* - g^*| = 2$. One such example is simple: let f^* be uniform on $[0,1]^2 \cup [1,2]^2$ and let g^* be uniform on

$[0,1] \times [1,2] \cup [1,2] \times [0,1]$.

The inequality of Theorem 1 seems to suggest that one cannot lose by first estimating f^* by g^*, and then taking the marginal density g to estimate the marginal density f. What one should remember though is that $\int |f^* - g^*|$ is usually very large to start with: if g^* is a nonparametric estimate of f^*, the rate of convergence to 0 of $\int |f^* - g^*|$ is normally a function of d.

2. COMPOSITION (MIXTURES) OF DENSITIES

When two densities f and g on R^d can be written as finite mixtures $\sum p_i f_i$ and $\sum p_i g_i$, where the f_i's and g_i's are densities on R^d, and $(p_1, \ldots, p_i, \ldots)$ is a probability vector, then we have the following theorem:

THEOREM 2.

$$\int \left| \sum p_i f_i - \sum p_i g_i \right| \leq \sum p_i \int |f_i - g_i|.$$

The inequality of Theorem 2 is trivial, yet it has interesting implications. Assume, for example, that we know that $f = pf_1 + (1-p)f_2$, where p is known (this is only for the sake of simplicity), f_1 is known to belong to some small parametric family (e.g., the family of normal densities), and f_2 is unknown. If someone shows us samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ drawn from f_1 and f_2, respectively, we could form an estimate of f by properly combining a parametric estimate of f_1 with a nonparametric estimate of f_2. Since the error of the last estimate usually dominates, we see that the total error is approximately bounded by $(1-p)$ times this error for sample size m. On the other hand, had we disregarded the f_1 information and lumped the samples together in a nonparametric estimate of f, we would have obtained a nonparametric estimate of sample size $n+m$. If m/n is about $(1-p)/p$, this is usually bigger than the former type of error.

When a sample from f is available and f_1 is partially or completely known, one should be able to take advantage of the additional information, although it is not immediately obvious how one could do so.

Finally, we note that the inequality of Theorem 2 can be grossly inadequate. For example, if f_1 and g_2 are uniform on $[0,1]$ and f_2 and g_1 are uniform on $[1,2]$, while $p_1 = p_2 = \frac{1}{2}$, we have $\int |f - g| = 0$ and $\int |f_i - g_i| = 2$, all i.

3. RESTRICTIONS OF DENSITIES

Consider a density f and an estimate g, also a density. Sometimes we know that f has support A. To avoid nonsensical situations (e.g., a density estimate that puts some mass on the negative numbers, while the density to be estimated is the density of a positive random variable), we can replace g by its *restriction* to A, that is,

$$g^*(x) = \frac{g(x)I_A}{\int_A g}.$$

We encountered such a restriction in the development of the transformed kernel estimate (Chapter 9). The restriction is always a better estimate than the original estimate g:

THEOREM 3.

$$\int |f - g^*| \le \int |f - g|.$$

Proof.

$$\int |f - g^*| = 2\int_A \left(f - \frac{g}{\int_A g} \right)_+ \le 2\int_A (f - g)_+ \le 2\int (f - g)_+ = \int |f - g|.$$

4. NONNEGATIVE PROJECTIONS

We assume again that f and g are a density and its estimate on R^d. However, while $\int g = 1$, g could take negative values: for example, this was the case for Bartlett's estimate (Section 7.5). The function

$$g^* = \frac{gI_A}{\int_A g}, \qquad A = \{x: g(x) > 0\},$$

is a valid density, and we will call it the *nonnegative projection* of g. Again, it is always better than the original g:

THEOREM 4.

$$\int |f - g^*| \leq \int |f - g|.$$

Proof. Since $\int g = 1$, we have $\int_A g \geq 1$. Therefore, $g \geq g^*$ on A.Thus,

$$\int |f - g^*| = 2\int (g^* - f)_+$$

$$= 2\int_A (g^* - f)_+ + 2\int_{A^c} (g^* - f)_+ = 2\int_A (g^* - f)_+$$

$$\leq 2\int_A (g - f)_+ = 2\int_{g>f} (g - f)_+ = \int |g - f|.$$

5. PRODUCT DENSITIES

The question posed here is the following: when univariate densities $f_i, 1 \leq i \leq d$, are estimated by univariate densities g_i, how well does $\prod_{i=1}^d g_i(x_i)$ estimate $\prod_{i=1}^d f_i(x_i), x_i \in R$? We offer the following inequalities:

THEOREM 5. *Let H_{pi} be the Hellinger distance between f_i and g_i, that is,*

$$H_{pi} = \left(\int |f_i^{1/p} - g_i^{1/p}|^p \right)^{1/p}, \qquad 1 \leq i \leq d,$$

and define

$$L_i = \int f_i \log\left(\frac{f_i}{g_i}\right), \qquad 1 \leq i \leq d.$$

Then, recalling that

$$\int \min\left(\prod f_i, \prod g_i\right) = 1 - \frac{1}{2}\int \left|\prod f_i - \prod g_i\right|,$$

we have

$$\int \min\left(\prod f_i, \prod g_i\right) \leq \exp\left(-\sum \frac{1}{2} H_{2i}^2\right) \leq \exp\left(-\sum \frac{1}{8} H_{1i}^2\right),$$

and

$$\int \min\left(\prod f_i, \prod g_i\right) \geq \exp\left(-\sum \frac{H_{1i}}{2 - H_{1i}}\right),$$

$$\int \min\left(\prod f_i, \prod g_i\right) \geq \tfrac{1}{2}\exp\left(-\sum L_i\right).$$

Proof. The first inequality follows from

$$\int \min\left(\prod f_i, \int g_i\right) \leq \prod_i \int \sqrt{f_i g_i} = \prod_i \left(1 - \frac{1}{2}H_{2i}^2\right) \leq \exp\left(-\sum \frac{1}{2}H_{2i}^2\right)$$

and the inequality $H_{2i} \geq \frac{1}{2}H_{1i}$ (Theorem 8.4). The second inequality follows from

$$\int \min\left(\prod f_i, \prod g_i\right) \geq \int \prod \min(f_i, g_i) = \prod \int \min(f_i, g_i)$$

$$= \prod \left(1 - \frac{1}{2}H_{1i}\right) \geq \exp\left(-\sum \frac{H_{1i}}{2 - H_{1i}}\right).$$

Finally, the last inequality follows directly from Theorem 8.2.

If $f_i = f$, $g_i = g$ and $\int |f - g| \neq 0$, we see that $\int |\prod f_i - \prod g_i| \to 2$ as $d \to \infty$. This is why the inequalities in Theorem 5 were formulated in terms of $\int \min(\prod f_i, \prod g_i)$.

6. RADIALLY SYMMETRIC DENSITIES

We say that f^* is a *radially symmetric density* on R^d when it is the density of a random variable YZ, where Y is a random variable with some density f on $[0, \infty)$, and Z is independent of Y and uniformly distributed on the surface of the unit hypersphere in R^d (thus, $\|Z\| = 1$ with probability 1). If g^* is another radially symmetric density associated with a density g on $[0, \infty)$, then we have the following theorem:

THEOREM 6.

$$\int |f^* - g^*| = \int |f - g|.$$

Proof. Let B be a Borel set of R^d. Then, if μ is the uniform measure on the unit hypersphere of R^d, we have $P(YZ \in B) = \int f(x)\mu(B/x)\,dx$. Thus,

$$\begin{aligned}\int |f^* - g^*| &= 2 \sup_B \left| \int_B f^* - \int_B g^* \right| \\ &= 2 \sup_B \left| \int f(x)\mu\left(\frac{B}{x}\right) dx - \int g(x)\mu\left(\frac{B}{x}\right) dx \right| \\ &= 2 \sup_B \left| \int_{f>g} (f-g)\mu\left(\frac{B}{x}\right) dx - \int_{f<g} (g-f)\mu\left(\frac{B}{x}\right) dx \right| \\ &\le 2 \max\left(\int_{f>g} (f-g), \int_{f<g} (g-f) \right) = \int |f-g|.\end{aligned}$$

Also, by taking $B = \{x\colon x \in R^d, f(\|x\|) > g(\|x\|)\}$, we see that the inverse inequality must be true too.

The importance of Theorem 6 is that all our one-dimensional results for density estimation carry over to the problem of estimating radially symmetric densities in R^d with known center. In particular, if the radial symmetry is given, one should always try to estimate f, the density of Y, and not f^*. An estimate of f^* can always be obtained by reconstructing the radial symmetry from the univariate estimate g of f. It goes without saying that this is just one example of many situations in which a priori knowledge about a problem can be used to reduce the dimensionality (and thus the difficulty).

7. CONVOLUTIONS

Let us now consider a situation in which we want to estimate the density of $Y_1 + \cdots + Y_d$, where the Y_i's are independent random variables with common unknown density f, and a sequence $X_1, \ldots, X_n$ of independent random variables with density f is available. In most interesting cases, d is either bounded or at least very small compared to n, for otherwise we would be better off applying a local central limit theorem. It seems that there are probably better ways of doing this than by merely estimating f^{*d} (the d-fold convolution of f) by g^{*d}, where g is a standard estimate of f. In any case, we have the following theorem:

THEOREM 7.

$$\int |f^{*d} - g^{*d}| \le d \int |f - g|.$$

Proof. The inequality follows from the fact that for any sequence of four densities f, g, $\tilde{f}$, and $\tilde{g}$:

$$\int |f * \tilde{f} - g * \tilde{g}| \le \int |f * (\tilde{f} - \tilde{g})| + \int |(f - g) * \tilde{g}|$$

$$\le \int |f - g| + \int |\tilde{f} - \tilde{g}|.$$

Thus the L_1 error is not guaranteed to remain at the same value. On the other hand, the inequality used in the proof of Theorem 7 is very loose: just consider four gamma densities with unequal parameters a, b, c, d, but $a + c = b + d$.

When d is large compared to n, local limit theorems will play an increasingly important role. For example, if density f has mean μ, variance σ^2, and third central moment α, then, if g is normal $(d\mu, d\sigma^2)$,

$$\int |f^{*d} - g| = \frac{|\alpha|}{3\sqrt{2\pi d}}(1 + 4e^{-3/2}) + o\left(\frac{1}{\sqrt{d}}\right)$$

(see, e.g, Petrov, 1975, p. 213 or Sirazdinov and Mamatov, 1962).

8. UNIMODAL DENSITIES

Consider a unimodal density f on R with a mode at 0. *Khinchine's theorem* (see Feller, 1971, p. 158) states that there exists a distribution function F such that

$$f(x) = \int_x^\infty \frac{1}{y}\, dF(y); \quad f(-x) = \int_{-\infty}^{-x} -\frac{1}{y}\, dF(y), \qquad x > 0.$$

Thus, when f and g are two unimodal densities on $[0, \infty)$ with mode at 0 and corresponding Khinchine distribution functions F and G,

$$\int |f - g| = \int \left| \int_x^\infty \frac{dF(y) - dG(y)}{y} \right| dx \le \int \int_x^\infty \frac{|dF(y) - dG(y)|}{y}\, dx$$

$$\le \int |dF(y) - dG(y)|.$$

If F and G have densities f^* and g^*, then the following can be concluded:

THEOREM 8.

$$\int |f - g| \leq \int |f^* - g^*|.$$

The relation between f and f^* is: $f(x) = \int_x^\infty f^*(y)/y\, dy$, and, clearly, g is closer to f than g^* is to f^* for any pair of unimodal densities. Unfortunately, we do not ordinarily have a sample from f^* at our disposal.

9. APPLICATIONS IN DETECTION

We will pose the detection problem in one of its simplest forms: f and g are two known densities on R^d, and we are given a sample $X_1, \ldots, X_n$ of independent identically distributed random vectors known to have density f or density g. We are asked to decide between the two alternatives. More formally, we will let a number Z be 1 or 2 according to whether the common density is f or g. Then, our decision Y, a Borel measurable function of $X_1, \ldots, X_n$, is 1 or 2. We are interested in the indicator of error

$$L_n = I_{[Y \neq Z]}.$$

In classical detection theory, the problem is actually asymmetric: one is interested in $I_{[Y \neq 1]}$ and $I_{[Y \neq 2]}$, when $Z = 2$ and $Z = 1$, respectively, but one kind of error is worse than the other. This will not be considered here. Nor will we consider problems in which we have to decide between f and "not f." These are better covered in texts on goodness-of-fit tests. For additional information about the detection problem, one can consult Rao (1973, Section 7a). For example, Theorem 9 below is strongly related to Section 7a3 of Rao (1973). From now on we take Z random with $P(Z = 1) = p \in [0, 1]$. All the bounds and claims of Theorems 9–11 are valid for *all* p, including $p = 0$, $p = 1$. Note that $E(L_n) = pP_1(Y \neq 1) + (1 - p)P_2(Y \neq 2)$, where P_1, P_2 are the probabilities for $X_1, \ldots, X_n$ conditioned on $Z = 1$, $Z = 2$, respectively. When $p = \frac{1}{2}$, $E(L_n)$ is minimized by setting

$$Y = \begin{cases} 1 & \text{if } \prod_{i=1}^{n} f(X_i)/g(X_i) > 1, \\ 2 & \text{otherwise.} \end{cases}$$

This will be called the *optimal detector* or *maximum likelihood detector*. For

convenience, we can rewrite it as

$$Y = \begin{cases} 1 & \text{if } (1/n)\sum_{i=1}^{n} \log(f(X_i)/g(X_i)) > c, \\ 2 & \text{otherwise,} \end{cases}$$

where we will let c take any real value for the time being. We should note here that the optimal detector depends upon the ratio f/g only, and is thus invariant to monotone transformations of the coordinate axes, a property that is desirable for all detectors. The value of the sum in the definition of the optimal detector could be $+\infty$ or $-\infty$, but is always well-defined in view of $P(f(X_i) = g(X_i) = 0 \text{ for some } i) = 0$.

THEOREM 9. *If* $c \in (-\int g \log(g/f), \int f \log(f/g))$, *then* $L_n \to 0$ *almost surely as* $n \to \infty$ *for all* $f \neq g$ *(i.e.,* $\int |f - g| > 0$*). Note that the interval for* c *may be left infinite or right infinite or both. It always includes the value* 0. *For* $c = 0$, *we have, with* $q = \min(p, 1-p)$,

$$P(Y \neq Z) = E(L_n) \begin{cases} \geq \frac{q}{2} \exp\left(-n \min\left(\int g \log\left(\frac{g}{f}\right), \int f \log\left(\frac{f}{g}\right)\right)\right) \\ \geq q \exp\left(-n \int |f-g| / \left(2 - \int |f-g|\right)\right) \\ \leq \exp\left(-(n/8)\left(\int |f-g|\right)^2\right). \end{cases}$$

Proof. The three inequalities follow from the observation that

$$E(L_n) = p \int_{\pi g_i < \pi f_i} \pi g_i + (1-p) \int_{\pi f_i \leq \pi g_i} \pi f_i,$$

where $f_i = f(x_i)$, $g_i = g(x_i)$, and the integral is with respect to $dx_1\, dx_2 \cdots dx_n$. Thus, $E(L_n)/\int \min(\pi f_i, \pi g_i) \in (q, 1]$. Now apply Theorem 5.

Let us next recall that $\int f \log(f/g)$ takes values in $(0, \infty]$ when $f \neq g$ (Theorem 8.2). When we split the integral into its positive and negative parts by splitting log into $\log_+ + \log_-$, we notice that the negative part has a bounded contribution since

$$0 \geq \int f \log_-\left(\frac{f}{g}\right) \geq -\frac{1}{e}.$$

To see this, we use the fact that $\log_+(g/f) \le g/ef$, and thus that

$$0 \le \int f \log_+\left(\frac{g}{f}\right) \le \int \frac{g}{e} = \frac{1}{e}.$$

We will also need the following form of the strong law of large numbers: if $Z_1, \ldots, Z_n, \ldots$ are independent identically distributed random variables with $E(Z_{1-}) > -\infty$, we have $(1/n)\sum_{i=1}^n Z_i \to E(Z_1)$ almost surely as $n \to \infty$, even if $E(Z_1) = \infty$.

From this, we conclude that on $Z = 1$,

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{f(X_i)}{g(X_i)}\right) \to \int f \log\left(\frac{f}{g}\right)$$

almost surely as $n \to \infty$. Thus, when $c < \int f \log(f/g)$, we have $P(Z = 1, Y = 2 | X_1, \ldots, X_n) \to 0$ almost surely as $n \to \infty$. The remainder of the theorem follows by symmetry.

Other detectors can be used too. In view of their suboptimality, they should only be used in special circumstances. For example, when f and/or g are not exactly known, but are merely good sample-based density estimates, then the optimal detector could yield disastrous results because of its sensitivity to the events $f(X_i) = 0$ and $g(X_i) = 0$. In other words, in such situations we would like to have more robust detectors. We will introduce and discuss two such detectors here, the PRD (pattern recognition based detector) and the L1D (L_1 error based detector).

The PRD is defined as follows:

$$Y = \begin{cases} 1 & \text{if } (1/n)\displaystyle\sum_{i=1}^{n} I_{[f(X_i)/g(X_i)>1]} - I_{[f(X_i)/g(X_i)<1]} > c. \\ 2 & \text{otherwise.} \end{cases}$$

In fact, in the PRD, we are summing the individually optimal decisions. It is worthwhile to observe that there are cases in which the choice $c = 0$ yields a detector with probability of error $L_n \to 1$ almost surely when $Z = 1$. For example, let g be uniform on $[0, 1]$, and let f be 2 on $[0, a]$, and $(1 - 2a)/(1 - a)$ on $(a, 1]$, for $0 < a < \frac{1}{4}$. Despite the fact that $Z = 1$, we have $g(X_i) > f(X_i)$ with probability $1 - 2a > \frac{1}{2}$.

THEOREM 10. *If* $c \in (\int_{f>g} g - \int_{f<g} g, \int_{f>g} f - \int_{f<g} f)$ *(an interval that does not have to contain* 0*), then* $L_n \to 0$ *almost surely as* $n \to \infty$ *for all*

$f \neq g$. *If we take*

$$c = \frac{1}{2}\left(\int_{f>g}(f+g) - \int_{f<g}(f+g)\right),$$

then

$$E(L_n) \leq \exp\left(-\frac{n}{8}\left(\int|f-g|\right)^2\right).$$

Proof. Let us take $p = 1$ without loss of generality. The sum in the definition of Y for the PRD has independent identically distributed $[-1, 1]$-valued summands with mean $\int_{f>g} f - \int_{f<g} f$. The first part of the theorem follows by the strong law of large numbers and the symmetry of the problem.

Also, if c is taken as indicated, then

$$E(L_n) = P\left(\frac{1}{n}\sum_{i=1}^{n}(W_i - E(W_i)) < c - E(W_1)\right),$$

where $W_i = I_{[f(X_i)/g(X_i)>1]} - I_{[f(X_i)/g(X_i)<1]}$, and $E(W_1) = \int_{f>g} f - \int_{f<g} f$. We verify that

$$c - E(W_1) = \frac{1}{2}\left(\int_{f>g}(g-f) - \int_{f<g}(g-f)\right) = -\frac{1}{2}\int|f-g|.$$

Since $\{W_i - E(W_i)\}$ are independent zero mean $[-1 - E(W_1), 1 - E(W_1)]$-valued random variables, we obtain, by Hoeffding's inequality (Hoeffding, 1963),

$$E(L_n) \leq \exp\left(-2n\left(\tfrac{1}{2}\int|f-g|\right)^2/4\right) = \exp\left(-\frac{n}{8}\left(\int|f-g|\right)^2\right).$$

The L1D is defined by

$$Y = \begin{cases} 1 & \text{if } \frac{1}{n}\sum_{i=1}^{n}\left(\left(1 - \frac{g(X_i)}{f(X_i)}\right)_+ - \left(1 - \frac{f(X_i)}{g(X_i)}\right)_+\right) > c, \\ 2 & \text{otherwise,} \end{cases}$$

where $c \in R$ is a constant. We have a behavior very similar to that of the PRD:

THEOREM 11. *If* $c \in (-\int f(1 - g/f)_+^2, \int g(1 - f/g)_+^2$, *then* $L_n \to 0$ *almost surely as* $n \to \infty$ *for all* $f \neq g$. *If we take*

$$c = \frac{1}{2}\left(\int g\left(1 - \frac{f}{g}\right)_+^2 - \int f\left(1 - \frac{g}{f}\right)_+^2 \right),$$

then

$$E(L_n) \leq \exp\left(- \frac{n}{24}\left(\int |f - g| \right)^2 \right).$$

Proof. The summands in the definition of Y are random variables with means

$$\int f\left(1 - \frac{g}{f}\right)_+ - \int f\left(1 - \frac{f}{g}\right)_+ = \int g\left(1 - \frac{f}{g}\right)_+^2 \quad \text{(when } P = 1\text{)}$$

and

$$\int g\left(1 - \frac{g}{f}\right)_+ - \int g\left(1 - \frac{f}{g}\right)_+ = - \int f\left(1 - \frac{g}{f}\right)_+^2 \quad \text{(when } P = 2\text{)},$$

so that the first statement follows simply by the strong law of large numbers for independent bounded random variables.

For the inequality we will assume without loss of generality that $P = 1$. We note that

$$E(L_n) = P\left(\frac{1}{n} \sum_{i=1}^{n} (W_i - E(W_i)) < - \frac{\epsilon}{2} \right),$$

where W_i is the ith summand in the definition of Y, and

$$\epsilon = \int g\left(1 - \frac{f}{g}\right)_+^2 + \int f\left(1 - \frac{g}{f}\right)_+^2.$$

Thus, by a variation of Bennett's inequality (Bennett, 1962),

$$E(L_n) \leq \exp\left(- \frac{n\epsilon^2}{8(\sigma^2 + \epsilon/2)} \right),$$

where $\sigma^2 = \mathrm{Var}(W_1)$. The inequality is only valid for independent sum-

mands W_i with $|W_i| \le 1$ (which is the case here). But

$$\sigma^2 \le E(W_1^2) = \int f\left(1 - \frac{g}{f}\right)_+^2 + \int f\left(1 - \frac{f}{g}\right)_+^2 \le \varepsilon,$$

so that the inequality becomes $E(L_n) \le \exp(-n\varepsilon/12)$. But by Cauchy's inequality,

$$\varepsilon \ge \left(\int g\left(1 - \frac{f}{g}\right)_+\right)^2 + \left(\int f\left(1 - \frac{g}{f}\right)_+\right)^2 = \frac{1}{2}\left(\int |f - g|\right)^2,$$

and we are done.

There are countless other examples of detectors. For example, $L_n \to 0$ almost surely for the detector

$$Y = \begin{cases} 1 & \text{if } \frac{1}{n}\sum_{i=1}^{n}\left(\frac{f(X_i)}{g(X_i)} - \frac{g(X_i)}{f(X_i)}\right) > c, \\ 2 & \text{otherwise,} \end{cases}$$

when $c \in (-\int g^2/f + 1, \int f^2/g - 1)$, an interval that contains the interval $(-(\int |f - g|)^2, (\int |f - g|)^2)$ when $f \ne g$ (Theorem 8.3). The unbounded summands are very sensitive to differences in the tails and in the support of f and g, but this sensitivity will cause some problems when f and/or g are not exactly known.

In practice, f and g are seldom known. Frequently, we are given random vectors $Y_1, \ldots, Y_k$ and $Z_1, \ldots, Z_k$ with densities f and g, respectively. First, f and g are estimated by f_k and g_k, where f_k and g_k are densities on R^d. Then we are asked to classify $X_1, \ldots, X_n$ as having common density f or common density g. The indicator of error now is

$$L_n = I_{[Y \ne Z]}.$$

In most communication theoretic and information theoretic applications, k can be thought of as fixed, but n is sometimes flexible, as $X_1, \ldots, X_n$ can be considered as samples of a signal evolving in time. Thus, the quantity

$$L_k^* = \limsup_{n \to \infty} L_n \quad \text{(in the a.s. sense)}$$

captures very well how good f_k and g_k are. Obviously, since L_n can only take the values 0 and 1, L_k^* can only take the values 0 and 1. We cannot

realistically expect that $P(L_k^* = 1) = 0$. However, we will say that our detector is *consistent* if

$$\lim_{k \to \infty} P(L_k^* = 1) = 0.$$

Let us now try to derive the consistency for some large classes of detectors. It is left as a simple exercise to extend all that follows to the notion of strong consistency, that is, $L_k^* \to 0$ almost surely as $k \to \infty$. The detectors we will consider are based upon the existence of a function $H:[0, \infty] \to [-1, 1]$ with the property that for all $f \neq g$ (i.e., $\int |f - g| > 0$),

$$\int f H\left(\frac{f}{g}\right) > \int g H\left(\frac{f}{g}\right).$$

Examples of such functions H are $H(u) = I_{[u>1]} - I_{[u<1]}$ (which leads to the PRD and $H(u) = (1 - 1/u)_+ - (1 - u)_+$ (which leads to the L1D).

The detector is constructed as follows:

1. Construct the density estimates f_k and g_k from the samples $Y_1, \ldots, Y_k$ and $Z_1, \ldots, Z_k$, respectively. Compute the detector's threshold $c_k = \int \frac{1}{2}(f_k + g_k) H(f_k/g_k)$.
2. The decision Y is defined by

$$Y = \begin{cases} 1 & \text{if } \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} H(f_k(X_i)/g_k(X_i)) > c_k \\ 2 & \text{otherwise.} \end{cases}$$

THEOREM 12. *The sample-based detector described above is consistent when*

(i) *$f_k \to f$ and $g_k \to g$ in probability as $k \to \infty$, for almost all x;*
(ii) *H is continuous (this is satisfied for the L1D).*

Proof. Assume without loss of generality that $P = 1$. Because H is a bounded function, we see immediately that L_k^* is almost surely equal to

$$I_{[\int f H(f_k/g_k) \leq c_k]}.$$

By our definition of H, we know that

$$\int f H\left(\frac{f}{g}\right) > c = \int \frac{f+g}{2} H\left(\frac{f}{g}\right).$$

Thus, we need only show that $c_k \to c$ in probability and that $\int fH(f_k/g_k) \to \int fH(f/g)$ in probability as $k \to \infty$. The latter fact follows directly from the Lebesgue dominated convergence theorem and assumptions (i) and (ii). Also,

$$|c_k - c| \le \frac{1}{2}\left(\int |f_k - f| + \int |g_k - g|\right) + \int \frac{f+g}{2}\left|H\left(\frac{f_k}{g_k}\right) - H\left(\frac{f}{g}\right)\right|$$

tends to 0 in probability by (i), Glick's Theorem 2.8, and (ii).

Extensions of the maximum likelihood detector are not straightforward. One of the main obstacles is the resolution of the problem that for some i's $\log(f_k(X_i)/g_k(X_i))$ takes the value $-\infty$, while for some other i's it takes the value $+\infty$. This instability can be resolved of course by truncating the logarithm from below and from above, a process called winsorization (see Huber (1981) for a discussion of winsorized maximum likelihood detectors). For the winsorized maximum likelihood detector, we clearly have a function H that satisfies the conditions of Theorem 12 and the consistency follows without work.

10. SYMMETRIZATION AND PERMUTATION INVARIANCE.

In this section, f is an arbitrary density on R^d, and unless specified otherwise, f_n is an arbitrary density estimate based on a sample of size n drawn from f. It is well-known that estimates should improve if more observations are taken, and that estimates that are not symmetric functions of the data can be improved by symmetrization. Yet it is another matter to show these improvements quantitatively for a fixed n. This will be done here. At the same time we take the opportunity to illustrate the beauty and elegance of the theory of Schur-convexity by proving all our results starting from an inequality of Marshall and Proschan (1965).

LEMMA 1 (Marshall and Proschan, 1965). *Let ϕ be a convex function of its n arguments, and let it be symmetric in its arguments. Let $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_n)$ be two weight vectors such that a majorizes b ($a \succ b$), that is,*

$$\sum_{i=1}^{k} a_{[i]} \ge \sum_{i=1}^{k} b_{[i]}, \qquad k = 1, \ldots, n,$$

with equality for $k = n$, where $a_{[1]} \geq \cdots \geq a_{[n]}$ and $b_{[1]} \geq \cdots \geq b_{[n]}$ are reorderings of a and b. If $X_1, \ldots, X_n$ are random vectors with permutation invariant distributions, then

$$E(\phi(a_1 X_1, \ldots, a_n X_n)) \geq E(\phi(b_1 X_1, \ldots, b_n X_n)).$$

THEOREM 13. *Let f_n be an estimate of the form*

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_n(x, X_i),$$

where the functions K_n are arbitrary measurable functions. If $X_1, \ldots, X_n$, $X_{n+1}, \ldots, X_{n+m}$ are independent identically distributed random variables, then

$$E\left(\int |g_n - f|\right) \leq E\left(\int |f_n - f|\right),$$

where

$$g_n(x) = \frac{1}{n+m} \sum_{i=1}^{n+m} K_n(x, X_i).$$

REMARK. For the kernel estimate (Parzen, 1962; Rosenblatt, 1956), with fixed kernel K and fixed smoothing factor h, the L_1 error is a decreasing function of n.

The proof of Theorem 13 will be postponed until after the presentation of the results.

THEOREM 14. *Let f_n be an estimate of the form*

$$f_n(x) = \sum_{i=1}^{n} w_{ni} K_n(x, X_i),$$

where the w_{ni}'s are weights summing to 1, *and K_n is as in Theorem* 13. *Then*

$$E\left(\int |g_n - f|\right) \leq E\left(\int |f_n - f|\right),$$

where

$$g_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_n(x, X_i).$$

and note that this is a convex symmetric function. Thus, by Lemma 1, and the fact that $a = (1, 0, 0, \ldots, 0) \succ b = (1/n!, \ldots, 1/n!)$, we have

$$E\left(\left|\sum_{\sigma=1}^{n!} a_\sigma Y_\sigma\right|\right) = E(|Y_1|) = E(|f_n(x, X_1, \ldots, X_n) - f(x)|)$$

$$\geq E\left(\left|\sum_{\sigma=1}^{n!} \frac{1}{n!} Y_\sigma\right|\right) = E(|g_n(x, X_1, \ldots, X_n) - f(x)|).$$

Taking the integral on left and right with respect to dx gives us Theorem 16. (We note in passing that $E(|g_n - f|^p) \leq E(|f_n - f|^p)$, all x, all $p \geq 1$.)

Theorem 15 follows from Theorem 16 without further work. To prove Theorem 13, we construct g_n as in Theorem 16, and note that

$$g_n(x, X_1, \ldots, X_{n+m}) = \frac{1}{(n+m)!} \sum_\sigma \frac{1}{n} \sum_{i=1}^{n} K_n(x, X_{\sigma(i)})$$

$$= \frac{1}{(n+m)!} \frac{1}{n} \sum_{i=1}^{n+m} K_n(x, X_i) \frac{(n+m)!}{n+m} n$$

$$= \frac{1}{n+m} \sum_{i=1}^{n+m} K_n(x, X_i).$$

For Theorem 14, we can apply Lemma 1 directly with $a = (w_{n1}, \ldots, w_{nn})$, $b = (1/n, \ldots, 1/n)$, $\phi(u_1, \ldots, u_n) = |\sum_{i=1}^{n} u_i|$, and with X_i formally replaced by $K_n(x, X_i) - f(x)$.

REFERENCES

G. Bennett (1962). Probability inequalities for the sum of independent random variables, *Journal of the American Statistical Association* **57**, pp. 33–45.

W. Feller (1971). *An Introduction to Probability Theory and Its Applications*, Vol. 2, Wiley, New York.

W. Hoeffding (1963). Probability inequalities for the sum of bounded random variables, *Journal of the American Statistical Association* **58**, pp. 13–30.

P. J. Huber (1981). *Robust Statistics*, Wiley, New York.

A. W. Marshall and F. Proschan (1965). An inequality for convex functions involving majorization, *Journal of Mathematical Analysis and Applications* **12**, pp. 87–90.

E. Parzen (1962). On the estimation of a probability density and the mode, *Annals of Mathematical Statistics* **33**, pp. 1065–1076.

V. V. Petrov (1975). *Sums of Independent Random Variables*, Springer-Verlag, New York.

C. R. Rao (1973). *Linear Statistical Inference and Its Applications*, Wiley, New York.

M. Rosenblatt (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics* **27**, pp. 832–837.

S. H. Sirazdinov and M. Mamatov (1962). On convergence in the mean for densities, *Theory of Probability and Its Applications* **7**, pp. 424–428.

G. Walter and J. Blum (1979). Probability density estimation using delta sequences, *Annals of Statistics* **7**, pp. 328–340.

W. Wertz (1976). Invariant density estimation, *Monatshefte für Mathematik* **81**, pp. 315–324.

CHAPTER 12

Estimators Based on Orthogonal Series

1. DEFINITIONS

The rich theory of orthogonal functions (Sansone, 1977; Szegö, 1975) can be used in the design of density estimators. There are of course a few problems with such an extrapolation: the original mathematical framework tends to ignore the fact that the estimated function is a density, and most approximations of functions by partial sums of orthogonal series expansions are not densities because they are either not in L_1 or violate the positivity constraint.

We start with an *orthonomal system* defined on a set B, usually R or $[-\pi,\pi]$. The functions of the orthonormal system, $p_0, p_1, \ldots$ satisfy, by definition,

$$\int_B p_i p_j = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}.$$

For a function f on B, we define its *Fourier coefficients* a_i by

$$a_i = \int_B f p_i.$$

A function f on B may or may not have an *expansion* in terms of the p_i's, depending upon whether $\sum_{i=0}^{\infty} a_i p_i(x)$ converges and is equal to $f(x)$, or not. This series, if its exists, is called the *Fourier series of* f. Without the existence of the Fourier series, it is hopeless to reconstruct f by using orthogonal series. It is thus important to characterize those situations in which f has an expansion in terms of the p_i's.

An orthonormal system is called *complete* in $L_p(B)$ if for any function $f \in L_p(B)$, $\int f p_i = 0$, all i, implies $f = 0$ almost everywhere. The system is

called a *basis* in $L_p(B)$ if for every $f \in L_p(B)$ there is a unique convergent series expansion $\sum a_i p_i$. It is known that when B is compact, a system is complete on $L_2(B)$ if and only if

$$\int_B f^2 = \sum_{i=0}^{\infty} a_i^2, \quad \text{all } f \in L_2(B)$$

This is called *Bessel's equality*. In that case, we actually have convergence of the partial sums in $L_2(B)$:

$$\int_B \left(f - \sum_{i=0}^{m} a_i p_i \right)^2 \to 0 \quad \text{as } m \to \infty.$$

(See, e.g., Sansone, 1977, p. 23.) There is no L_1 analogue of this property. If we want to study the convergence of the *partial sums*

$$S_m(f) = \sum_{i=0}^{m} a_i p_i$$

(sometimes we will write $S_m(f, x)$ to make the dependence upon x explicit) in $L_1(B)$, then we cannot just use the Cauchy–Schwarz inequality to make the bridge to L_2 theory:

$$\int_B |S_m(f) - f| \le \sqrt{\lambda(B) \int_B (S_m(f) - f)^2}.$$

Indeed, this would force us to introduce the condition $f \in L_2(B)$. The condition $\lambda(B) < \infty$ is less restrictive since we can always transform the data montonically to a compact interval; see, for example, Chapter 9.

One of the exciting features of orthogonal series expansions is that if a function has a finite term expansion, then approximation and hopefully estimation is very easy. If we are given a sample $X_1, \ldots, X_n$ drawn from f, then $a_i = \int f p_i$ can be estimated without bias by

$$a_{ni} = \frac{1}{n} \sum_{j=1}^{n} p_i(X_j),$$

and $f(x)$ can be estimated by

$$f_n(x) = \sum_{i=0}^{m} a_{ni} p_i(x).$$

If $f(x) = \sum_{i=0}^{m} a_i p_i(x)$ for a finite m, then f_n is an unbiased estimate of f. In general, f does not have a finite term expansion, and we will be forced to let the parameter m tend to infinity with n. Thus, f_n, the *orthogonal series estimate*, can be viewed as a direct generalization of parametric estimates.

There is also another way of writing down f_n:

$$f_n(x) = \frac{1}{n} \sum_{j=1}^{n} K_m(x, X_j),$$

where

$$K_m(x, y) = \sum_{i=0}^{m} p_i(x) p_i(y)$$

is called the *kernel*. This form will be called the *regular form*, because it reminds us of the kernel estimate (see Section 8 below for more on the connection with the kernel estimate).

One relevant result from Fourier analysis is the *Christoffel–Darboux summation formula* for orthogonal polynomials (see, e.g., Szegö, 1975, pp. 42–44):

$$K_m(x, y) = b_m \frac{p_{m+1}(x) p_m(y) - p_m(x) p_{m+1}(y)}{x - y},$$

where $b_m = k_m / k_{m+1}$ and k_m is the coefficient of the highest power of x in the polynomial p_m.

The integer m can be considered as a smoothing factor. We will mainly be concerned in this chapter with the consistency and the rate of convergence of orthogonal series estimators when $m = m_n$ is a sequence of integers. In particular, we will illustrate how some systems are not rich enough to handle all densities on B, for example, trigonometric, Hermite, Laguerre, and Legendre systems. This, of course, should be weighed against a few formidable advantages of orthogonal series estimates, for example, their superb performance when f has a finite term expansion, or a rapidly convergent infinite expansion.

In density estimation, it is important to have estimates with $\int_B f_n = 1$. In an orthogonal series estimate on a compact set B, this can be insured by choosing the first function p_0 as follows:

$$p_0 = \frac{I_B}{\sqrt{\lambda(B)}},$$

where I is the indicator function, Clearly,

$$\int_B f_n = \sum_{i=0}^{m} a_{ni} \int_B p_i = \sum_{i=0}^{m} a_{ni} \sqrt{\lambda(B)} \int_B p_i p_0 = a_{n0} \sqrt{\lambda(B)}$$

$$= a_0 \sqrt{\lambda(B)} = \int_B f = 1.$$

The inclusion of the constant function in an orthonormal system on R will in general lead to estimates that are not functions of L_1.

Orthogonal series estimates were developed in a series of papers of Cencov (1962), Van Ryzin (1966), Schwartz (1967), Kronmal and Tarter (1968), Bosq (1969), Watson (1969) and Földes and Révész (1974). In this chapter, we will not deal with multivariate orthogonal series estimates. For the multivariate trigonometric series estimate, the reader could consult Kronmal (1968), Schüler (1976), Sterbuchner (1980), Stegbuchner (1981), Greblicki and Pawlak (1981), and Krzyzak and Pawlak (1982).

2. EXAMPLES OF ORTHONORMAL SYSTEMS

The *trigonometric system* on $B = [-\pi, \pi]$ is formed by the functions

$$p_0 = \frac{1}{\sqrt{2\pi}}, \quad p_{2i-1}(x) = \frac{\cos(ix)}{\sqrt{\pi}}, \quad p_{2i}(x) = \frac{\sin(ix)}{\sqrt{\pi}}, \qquad i \geq 1.$$

The corresponding orthogonal series estimate is called the *trigonometric series estimate* or the *Fourier series estimate* (see, e.g., Kronmal and Tarter, 1968). The trigonometric system is complete in $L_1[-\pi, \pi]$, but is not a basis for $L_1[-\pi, \pi]$. It will be convenient to write the trigonometric series estimate sometimes as

$$f_n(x) = \frac{1}{2\pi} + \sum_{i=1}^{m} \left(a_{n,2i-1} \frac{\cos(ix)}{\sqrt{\pi}} + a_{n,2i} \frac{\sin(ix)}{\sqrt{\pi}} \right)$$

and sometimes as

$$f_n(x) = \frac{1}{n} \sum_{j=1}^{n} D_m(x, X_j),$$

where D_m is the *Dirichlet kernel*

$$D_m(x, y) = \frac{\sin\left(\frac{2m+1}{2}(x-y)\right)}{2\pi \sin\left(\frac{x-y}{2}\right)}.$$

Note that the definition of m differs slightly from that given in Section 1 since we are in fact considering an expansion with $2m + 1$ terms.

The *Legendre polynomials* form an orthonormal system on $[-1, 1]$. There are many ways of defining these polynomials. For example, they can be defined by *Rodrigues' formula*

$$p_i(x) = \sqrt{\frac{2i+1}{2}}\,\frac{1}{2^i i!}\,\frac{d^i}{dx^i}\left((x^2-1)^i\right), \qquad i \geq 0.$$

The system is complete in $L_1[-1, 1]$ (Sansone, 1977, p. 191). The corresponding kernel $K_m(x, y)$ is

$$K_m(x, y) = \sum_{i=0}^{m} p_i(x)p_i(y)$$

$$= \frac{m+1}{\sqrt{2m+1}\sqrt{2m+3}}\,\frac{p_m(x)p_{m+1}(y) - p_{m+1}(x)p_m(y)}{y-x}.$$

Legendre series estimators were discussed by Crain (1974), Viollaz (1980), and Hall (1982). For various explicit forms of the p_i's and the derivation of K_m, for example, Sansone (1977) or Szegö (1975). There are several orthonormal systems that generalize the Legendre system, for example, Ferrer's functions (Sansone, 1977, pp. 246–253) and the Jacobi polynomials (Szegö, 1975, Chapter 4).

A function can be expanded sometimes into a Hermite series using the functions

$$p_i(x) = \frac{e^{x^2/2}}{\sqrt{2^i i!\sqrt{\pi}}}\,\frac{d^i}{dx^i}\left(e^{-x^2}\right), \qquad i \geq 0.$$

These functions form an orthonormal system, and are complete in $L_2(R)$. The *Hermite series estimate* was studied for use in density estimation by Schwartz (1967), Walter (1977), Bleuez and Bosq (1979), and Greblicki

(1981). Using (5.5.9) of Szegö (1975), one can derive the kernel

$$K_m(x, y) = \frac{m+1}{2} \frac{p_{m+1}(x)p_m(y) - p_m(x)p_{m+1}(y)}{x - y}.$$

The *Laguerre series estimate* on $B = [0, \infty)$ is based upon the orthonormal and complete (in $L_2[0, \infty)$) system of functions

$$p_i(x) = \left(\frac{\Gamma(i+1)}{\Gamma(i+\alpha+1)} x^{-\alpha} e^x \right)^{1/2} \cdot \frac{1}{i!} \frac{d^i}{dx^i}(x^{i+\alpha} e^{-x}), \qquad i \geq 0.$$

Here $\alpha > -1$ is a parameter of the system. For example, for $\alpha = 0$, we obtain

$$p_i(x) = e^{-x/2} \sum_{j=0}^{i} (-1)^j \frac{1}{j!} \binom{i}{j} x^j.$$

The kernel is

$$K_m(x, y) = \frac{\Gamma(m+2)}{\Gamma(m+\alpha+1)} \frac{p_{m+1}(x)p_m(y) - p_m(x)p_{m+1}(y)}{y - x}.$$

The *Haar orthonormal system* differs from all the previous systems in that it is a basis for all $L_p[0, 1]$. For a given integer m, the functions are defined as follows: define integers $k \geq 0$, and $j, 1 \leq j \leq 2^k$, by the equation $m = 2^k + j$. Then, set

$$p_m(x) = \begin{cases} 2^{k/2}, & x \in \left(\frac{j-1}{2^k}, \frac{j-1/2}{2^k} \right) \\ -2^{k/2}, & x \in \left(\frac{j-1/2}{2^k}, \frac{j}{2^k} \right) \\ 0, & \text{otherwise.} \end{cases}$$

This system has the desirable property that for all $f \in L_1[0, 1]$, $S_m(f) \to f$ almost everywhere, and $\int |S_m(f) - f| \to 0$. Its regular form nearly reduces to the histogram estimate (see, e.g., Bleuez and Bosq, 1979). In fact, its kernel takes only nonnegative values, so that it is easily seen that f_n itself is a density. The only difference with the histogram estimate with equal intervals is that the intervals for different values of m are properly nested (due to the dyadic construction of the functions p_m). The *Haar series*

estimate inherits all the properties of the histogram estimate, including the built-in limitation of an expected L_1 error rate that is bounded from below by a constant times $n^{-1/3}$, and including the consistency for all densities f on $[0, 1]$. It will not be treated elsewhere in this chapter.

3. GENERAL PROPERTIES

The following lemma gives us useful albeit crude upper and lower bounds for the expected L_1 error (see also Lemma 3.6):

LEMMA 1. *Let f_n be an orthogonal series estimate of f with $m = m_n$ terms, and let f have the formal series expansion (convergent or not)*

$$f(x) \sim \sum_{i=0}^{\infty} a_i p_i(x),$$

where the p_i's form an orthonormal system on a set B of R, and $a_i = \int_B f p_i$. Assume also that all p_i's are absolutely integrable on B. Then,

$$\begin{aligned} E\left(\int |f_n - f|\right) &\le \int |S_m(f) - f| + E\left(\int |f_n - E(f_n)|\right) \\ &\le \int |S_m(f) - f| + \int \sqrt{E\left((f_n - E(f_n))^2\right)} \\ &\le \int |S_m(f) - f| + \frac{1}{\sqrt{n}} \int \sqrt{E\left(K_m^2(x, X_1)\right)}\, dx; \end{aligned}$$

$$E\left(\int |f_n - f|\right) \ge \operatorname{Max}\left(\int |S_m(f) - f|, \frac{1}{2} \int E(|f_n - E(f_n)|)\right).$$

LEMMA 2. *Let $f \in L_2(B)$, and let $p_i, i \ge 0$, form an orthonormal system on B. Then, the orthogonal series estimate f_n with m terms has the following expected L_2 error (all integrals are over B):*

$$\begin{aligned} E\left(\int (f_n - f)^2\right) &= \frac{1}{n} \sum_{i=0}^{m} \left(\int f p_i^2 - a_i^2\right) + \sum_{i=m+1}^{\infty} a_i^2 \\ &\le \frac{1}{n} \sum_{i=0}^{m} \int f p_i^2 + \sum_{i=m+1}^{\infty} a_i^2 \\ &= \frac{1}{n} \int f(x) K_m(x, x)\, dx + \sum_{i=m+1}^{\infty} a_i^2. \end{aligned}$$

Proof. We begin with

$$\int (f_n - f)^2 = \int (f_n - E(f_n))^2 + \int (S_m(f) - f)^2.$$

Since Bessel's equality applies, the last term is equal to

$$\int \left(\sum_{i=m+1}^{\infty} a_i p_i \right)^2 = \sum_{i=m+1}^{\infty} a_i^2.$$

Here we used the orthonormality. Also,

$$\int (f_n - E(f_n))^2 = \int \left(\sum_{i=0}^{m} (a_i - a_{ni}) p_i \right)^2 = \sum_{i=0}^{m} (a_i - a_{ni})^2,$$

and the rest follows without work.

Because we require integrable estimates, it is only reasonable to require that all p_i's in our orthonormal system be absolutely integrable. However, this restriction has some ill side-effects when an orthogonal series estimate is considered on an unbounded set. The following lemma captures these side-effects.

LEMMA 3. *Let $p_i, i \geq 0$, be an orthonormal system on R (or on $[0, \infty)$), and let all p_i's be absolutely integrable. Then*

(i) *If f_n is translation-invariant (see Section 6.6), $\int f_n = 0$.*

(ii) *It is impossible that $\int f_n = 1$ almost surely for all f.*

Proof. If f_n is translation-invariant for all f, then $K_m(x, y)$ must be of the form $K_m^*(x - y)$ for some function K_m^*, and $\int K_m(x, y)\, dy$ should be independent of x. Thus,

$$p_i(x) \sum_{i=0}^{m} \int p_i(y)\, dy$$

should be independent of x. Because all p_i's are nondegenerate functions ($\int p_i^2 = 1$), and no p_i is equal to a constant almost everywhere (since $\int p_i^2 = 1$), we see that $\sum_{i=0}^{m} \int p_i(y)\, dy = 0$, and thus $\int K_m(x, y)\, dx = 0$, all y. Thus, $\int f_n = 0$.

To show (ii), we will argue by contradiction. If $\int f_n = 1$ almost surely, for all f, then $\sum_{i=0}^{m} \int p_i(x) p_i(y)\, dx = 1$ for almost all y. Squaring and integrat-

ing with respect to dy gives

$$\infty = \int \left(\sum_{i=0}^{m} p_i(y) \int p_i \right)^2 dy = \sum_{i=0}^{m} \int p_i^2 \left(\int p_i \right)^2 = \sum_{i=0}^{m} \left(\int p_i \right)^2,$$

which is clearly impossible by our assumptions.

Lemma 3 leaves no doubt as to the limitations of orthogonal series estimates on R or on $[0, \infty)$. As a result, we will mainly discuss the properties of orthogonal series estimates on compact sets.

4. THE TRIGONOMETRIC SERIES ESTIMATE: CONSISTENCY

The aim of this section is threefold. Densities on $[-\pi, \pi]$ will be constructed with a convergent Fourier series that cannot be estimated by the trigonometric series estimate, regardless of how large n is. To balance this, we will give weak sufficient conditions for consistency. Finally, we will address very briefly the issue of the necessity of these conditions.

THEOREM 1 (The Nonconsistency of the Trigonometric Series Estimate). *Let $a_m \downarrow 0$ be a sequence of numbers with $a_0 = 1/\sqrt{\pi}$, and convex considered as a function of m. Then*

$$\frac{1}{2\pi} + \sum_{i=1}^{\infty} a_i \frac{\cos(ix)}{\sqrt{\pi}}$$

exists (i.e., the series converges), and is the Fourier series of a density f on $[-\pi, \pi]$.

Let f_n be the trigonometric series estimate with parameter m and sample size n. If

$$\liminf_{m \to \infty} a_m \log m > \pi^{3/2}/2,$$

then

$$\inf_{n,m} E\left(\int |f_n - f| \right) > 0.$$

If $\lim_{n\to\infty} m = \infty$ *and* $\lim_{m\to\infty} a_m \log m = \infty$, *then*

$$\liminf_{n\to\infty} \frac{E\left(\int |f_n - f|\right)}{a_m \log m} \geq \frac{4}{\pi^{3/2}}.$$

Theorem 1 states that for many a density on $[-\pi, \pi]$, even densities with convergent Fourier series (except at one point), there is no possibility of estimating f, regardless of how n and m are chosen. The reason for this is a run-away bias ($\int |S_m(f) - f| \to \infty$ as $n \to \infty$, when $m \to \infty$, $a_m \log m \to \infty$). Thus, $E(\int |f_n - f|)$ can increase at any prescribed rate that is $o(\log m)$ when $m \to \infty$.

The proof requires several auxiliary results and in particular some properties of the *Dirichlet kernel* $D_m(x, y)$. Since this is a function of $x - y$ only, we take the liberty to write from now on

$$D_m(u) = \frac{\sin\left(\left(m + \frac{1}{2}\right)u\right)}{2\pi \sin(u/2)}.$$

Another important function is *Fejér's kernel*

$$F_m(u) = \frac{1}{m+1} \sum_{i=0}^{m} D_i(u) = \frac{1}{2\pi(m+1)} \left(\frac{\sin((m+1)u/2)}{\sin(u/2)} \right)^2.$$

LEMMA 4 (Properties of the Dirichlet and Fejér Kernels).

A. $\int_{-\pi}^{\pi} D_m(u)\, du = \int_{-\pi}^{\pi} F_m(u)\, du = 1$.

B. $|D_m(u)| \leq 1/2|u|, |u| \leq \pi; |D_m(u)| \leq (m+1)/2, |u| \leq \pi$.

C. $\int |D_m| \sim (4/\pi^2) \log m$ *as* $m \to \infty$. ($\int |D_m|$ *is called a Lebesgue constant.*)

D. $\int |D_m| \leq 2 + (4/\pi^2) \log m, m \geq 1$.

E. $F_m(u) \leq \pi/2(m+1)u^2, |u| \leq \pi; F_m(u) \leq (m+1)/4$.

Proof. Property A is well-known. It follows directly from the definitions of D_m and F_m and the orthonormality conditions. For property B, we have

$$|D_m(u)| \leq \left(2\pi \left|\sin\left(\frac{u}{2}\right)\right|\right)^{-1} \leq \sup_v \left|\frac{v}{\sin(v)}\right| \left|\frac{1}{\pi|u|}\right| \leq \frac{1}{2|u|}$$

and

$$|D_m(u)| \leq \sup_v \left|\frac{\sin v}{v}\right| \frac{(m+1/2)}{\pi} \frac{|u/2|}{\sin(|u/2|)} \leq \frac{1}{2}\left(m + \frac{1}{2}\right),$$

where all u, v take values in $[-\pi, \pi]$ only.

For property C, we refer the reader to Bary (1964, Vol. 1, pp. 108–109). One half of property C follows of course from property D, which will be shown here in full.

$$\int |D_m| = \frac{1}{\pi}\int_0^{\pi}\left|\frac{\sin\left(\left(m+\frac{1}{2}\right)u\right)}{\sin(u/2)}\right| du = \frac{2}{\pi}\int_0^{\pi/2}\left|\frac{\sin((2m+1)y)}{\sin(y)}\right| dy$$

$$\leq \frac{2}{\pi}\int_0^{\pi/2}\left|\frac{\sin((2m+1)y)}{y}\right| dy + \frac{2}{\pi}\int_0^{\pi/2}\left|\frac{1}{\sin(y)} - \frac{1}{y}\right| dy.$$

Let us call the latter two terms I_1 and I_2. We verify quickly that $I_2 \leq \pi^2/48$. This follows from

$$\left|\frac{y - \sin(y)}{y\sin(y)}\right| \leq \left|\frac{y^3/6}{(2/\pi)y^2}\right| \leq \frac{\pi y}{12}, \qquad 0 \leq y \leq \frac{\pi}{2},$$

and integrating this bound.

To treat I_1, we argue as in Bary (1964, Vol. 1, pp. 107–108) or Edwards (1979, pp. 80–81). We have, putting $(2m+1)y = t$,

$$I_1 = \frac{2}{\pi}\int_0^{(2m+1)\pi/2}\left|\frac{\sin(t)}{t}\right| dt$$

$$= \frac{2}{\pi}\sum_{k=0}^{2m}\int_{k\pi/2}^{(k+1)\pi/2}\frac{|\sin(t)|}{t}\, dt$$

$$\leq \frac{2}{\pi}\sum_{k=1}^{2m}\int_{k\pi/2}^{(k+1)\pi/2}\frac{|\sin(t)|}{k/2}\, dt + 1$$

$$= \frac{4}{\pi^2}\sum_{k=1}^{2m}\frac{1}{k} + 1$$

$$\leq \frac{4}{\pi^2}(1 + \log(2m)) + 1.$$

Property D now follows after noting that $\pi^2/48 + 1 + 4/\pi^2 + 4\log(2)/\pi^2 \leq 2$.

Finally, property E can be obtained as follows:

$$F_m(u) \le \left(2\pi(m+1)\sin^2\left(\frac{u}{2}\right)\right)^{-1} \le \left(2(m+1)u^2\right)^{-1}\pi;$$

and

$$F_m(u) \le \frac{1}{m+1}\sum_{i=0}^{m}\frac{1}{2}\left(m+\frac{1}{2}\right)$$

$$\le \frac{1}{2(m+1)}\left(\frac{m+1}{2}+\frac{m(m+1)}{2}\right) = \frac{m+1}{4}.$$

LEMMA 5 (The Fejér–Lebesgue Theorem; see, e.g., Bary, 1964, Vol. 1, pp. 139–140). *If we define*

$$\sigma_m(f) = \int_{-\pi}^{\pi} f(t)F_m(t-x)\,dt,$$

then for all densities f on $[-\pi,\pi]$, $\sigma_m(f) \to f$ as $m \to \infty$, almost all x, and $\int|\sigma_m(f)-f| \to 0$ as $m \to \infty$.

REMARK. It will be convenient to define f outside $[-\pi,\pi]$ by periodicity. This will make the notation simpler. Note also that D_m and F_m are periodic with period 2π.

Proof of Lemma 5.

$$|\sigma_m(f)-f| \le \int|f(u+x)-f(x)|F_m(u)\,du$$

$$\le \delta^{-1}\int_{|u|\le\delta}|f(u+x)-f(x)|\,du\left(\frac{m+1}{4}\right)\delta$$

$$+\int_{|u|\ge\delta}\frac{|f(u+x)-f(x)|\pi}{2(m+1)u^2}\,du.$$

Here $\delta > 0$ was arbitrary. Choose $\delta = 1/(1+m)$, and note that the first term on the right-hand side is $o(1)$ for almost all x by the Lebesgue density theorem. To treat the second term, we define $g(u) = |f(u+x)-f(x)|$ and $G(u) = \int_0^u g(v)\,dv$. Thus, $G(u) = o(u)$ as $u \downarrow 0$ for almost all x. By partial

integration, we observe that

$$\int_{\pi \geq u \geq \delta} g(u) \frac{1}{(m+1)u^2}\, du \leq \frac{1}{m}\left(\frac{G(\pi)}{\pi^2} - \frac{G(\delta)}{\delta^2} + \int_{\delta}^{\pi} \frac{2G(u)}{u^3}\, du \right)$$

$$= o(1) + \frac{2}{m} \int_{1/(1+m)}^{\pi} \frac{G(u)}{u^3}\, du$$

$$= o(1) + \frac{2}{m} o\left(\int_{1/m}^{\infty} u^{-2}\, du \right)$$

$$= o(1)$$

for almost all x. The second part of the lemma follows from the first part and the observation that σ_m is a density for all m.

Proof of Theorem 1. We start with some well-known properties of Fourier series. Consider the series

$$\frac{a_0}{\sqrt{2\pi}} + \sum_{i=1}^{\infty} a_i \frac{\cos(ix)}{\sqrt{\pi}}$$

when the sequence $a_0\sqrt{2}, a_1, a_2, a_3, \ldots$ is nonincreasing and tends to 0. This series converges everywhere on $[-\pi, \pi]$ except perhaps at 0, and the convergence is uniform on $[\varepsilon, \pi]$, all $\varepsilon > 0$ (see, e.g., Bary, 1964, Vol. 1, pp. 87–88, for a simple proof). If this sequence is also convex, then the series converges to a nonnegative integrable function f on $[-\pi, \pi]$ (except perhaps at $x = 0$), and is the Fourier series of this function, that is,

$$a_i = \int_{-\pi}^{\pi} f(x) \frac{\cos(ix)}{\sqrt{\pi}}\, dx, \qquad i \geq 1;$$

$$a_0 = \int_{-\pi}^{\pi} f(x) \frac{1}{\sqrt{2\pi}}\, dx.$$

(See Bary, 1964, Vol. 1, pp. 92–94.) Since a_0, thus defined, is $1/\sqrt{2\pi}$ in our case, f is in fact a density on $[-\pi, \pi]$.

For convenience in notation, we define $b_0 = a_0\sqrt{2/\pi}$, $b_i = a_i/\sqrt{\pi}$, $i \geq 1$, so that

$$S_m(f) = \frac{1}{2} b_0 + \sum_{i=1}^{m} b_i \cos(ix),$$

and b_i tends to 0 as $i \to \infty$ in a monotone manner, and the sequence is convex. Thus, $\Delta b_i = b_i - b_{i+1} \geq 0$, all i, and $\Delta^2 b_i = \Delta b_i - \Delta b_{i+1} \geq 0$, all i.

We note that $D_m(x) = \frac{1}{2} + \sum_{i=1}^{m} \cos(ix)$ and that $\sum_{i=0}^{m} D_i(x) = (m+1)F_m(x)$. Using these identities and Abel's transformation, we have

$$\begin{aligned}\frac{1}{\pi} S_m(f) &= \sum_{i=0}^{m-1} \Delta b_i D_i(x) + b_m D_m(x) \\ &= \Delta b_{m-1} \sum_{i=0}^{m-1} D_i(x) + \sum_{i=0}^{m-2} \Delta^2 b_i \sum_{k=0}^{i} D_k(x) + b_m D_m(x) \\ &= \Delta b_{m-1} m F_{m-1}(x) + \sum_{i=0}^{m-2} \Delta^2 b_i \cdot (i+1) F_i(x) + b_m D_m(x).\end{aligned}$$

The first two terms are nonnegative. (Incidentally, since the last term in the last expression is $o(1)$ for $x \neq 0$, we have shown that the partial sums converge to a nonnegative function.) The integrals of the first two terms over $[-\pi, \pi]$ total

$$m\Delta b_{m-1} + \sum_{i=0}^{m-2} (i+1)\Delta^2 b_i = \sum_{i=0}^{m-1} \Delta b_i = b_0 - b_m.$$

Therefore,

$$\begin{aligned}\int |S_m(f)| &\geq \left(b_m \int |D_m| + b_m - b_0 \right) \pi \\ &= \sqrt{\pi} \left(a_m \int |D_m| + a_m - a_0 \sqrt{2} \right) = a_m \sqrt{\pi} \left(1 + \int |D_m| \right) - 1,\end{aligned}$$

and thus, by Lemma 1,

$$E\left(\int |f_n - f| \right) \geq \int |S_m(f) - f| \geq a_m \sqrt{\pi} \left(1 + \int |D_m| \right) - 2.$$

By Lemma 4, this lower bound is $a_m \log(m)$. $(4/\pi^{3/2} + o(1)) - 2$. The last statement of Theorem 1 follows directly from this. For the second statement of Theorem 1, we note that there exist positive constants c, M such that

$$\inf_{m \geq M} \int |S_m(f) - f| \geq c.$$

But since obviously, $\inf_{m \leq M} \int |S_m(f) - f| > 0$ for all finite M, we are done.

The examples in Theorem 1 are all nice, because the Fourier series converges at all $x \neq 0$. It is perhaps worthy to note that there are $f \in L_1[-\pi, \pi]$ for which $S_m(f)$ does not converge at any point; these f are in a sense worse than those exhibited in Theorem 1. Since we are not directly interested in pointwise convergence, we will merely state some known results about the pointwise convergence of partial sums of integrable functions f.

LEMMA 6 (Pointwise Convergence of Fourier series).

A. *For all* $f \in L_1[-\pi, \pi]$, $S_{m_k}(f) \to f$ *almost everywhere for some subsequence* m_k. *Furthermore,* $S_m(f) = o(\log m)$ *for almost all* x, *and* $\int |S_m(f)| = o(\log m)$.

B. *If* $f \in L_p[-\pi, \pi]$ *for some* $p > 1$, *then* $S_m(f) \to f$ *for almost all* x.

C. *There exists an* f *in* $L_1[-\pi, \pi]$ *such that* $\limsup_{m \to \infty} S_m(f) = \infty$, *all* x. *In fact, for some* f *in* $L_1[-\pi, \pi]$, $\limsup_{m \to \infty} |S_m(f)|/\log\log m = \infty$, *all* x.

D. *For all sequences* $m_k \uparrow \infty$, *there exists an* f *in* $L_1[-\pi, \pi]$ *such that* $\limsup_{k \to \infty} S_{m_k}(f) = \infty$, *almost all* x.

E. *For all sequences* $c_m \downarrow 0$, *there exists an* f *in* $L_1[-\pi, \pi]$ *such that* $\limsup_{m \to \infty} |S_m(f)|/(c_m \log m) = \infty$.

REMARK. Part A can be found in Zygmund (1959, Section 7.3) and Edwards (1979, pp. 167, 180). For $p = 2$, property B is known as Carleson's theorem (Carleson, 1966). The general statement for $p > 1$ was proved by Hunt (1968) (see also Mozzochi (1971) and the book by Jorsboe and Mejlbro (1982) for other proofs of the profound *Carleson–Hunt theorem*). The first part of C is known as *Kolmogorov's counterexample* (Kolmogorov, 1926; see Zygmund, 1959, Section 8.4). The second part of C concerns a sharpening due to Körner (1981), who also proved D, based on ideas of Kahane and Stein. For property E, one can consult Edwards (1979, p. 180). We note also that the Carleson–Hunt theorem was generalized to d dimensions by Fefferman (1971) and Sjölin (1971).

After having established the lack of universal consistency of the trigonometric series estimate, it is possible nevertheless to say something positive for certain classes of densities. We will not study the strong convergence of the L_1 error of the estimate. The following lemma will help us to handle the bias term $\int |S_m(f) - f|$.

LEMMA 7. *Let* f *be a density on* $[-\pi, \pi]$. *Then*

(i) $\lim_{m \to \infty} \int |S_m(f) - f|^p = 0$, *all* $p \in (0, 1)$.

(ii) $\lim_{m \to \infty} \int |S_m(f) - f| = 0$ *whenever* $f \in L_p$, *some* $p > 1$.

(iii) $\lim_{m \to \infty} \int |S_m(f) - f| = 0$ *whenever* $\int f \log_+ f < \infty$, *and in fact* $\int |S_m(f)| \leq A \int f \log_+ f + B$ *for some universal constants* A, B.

REMARK. For Lemma 7, we refer to Section 7.3 of Zygmund (1959).

LEMMA 8. *For all densities f on $[-\pi, \pi]$, we have for the trigonometric series estimate with parameter m:*

$$E\left(\int |f_n - f|\right) \leq \sqrt{\frac{2m+1}{n}} + \int |S_m(f) - f|.$$

If $\lim_{n\to\infty} m = \infty$, *then*

$$E\left(\int |f_n - f|\right) \leq \sqrt{\frac{m}{\pi n}} \left(\int \sqrt{f} + o(1)\right) + \int |S_m(f) - f|.$$

THEOREM 2 (Consistency of the Trigonometric Series Estimate). *Let f be a density on $[-\pi, \pi]$ satisfying the peakedness condition $\int f \log_+ f < \infty$ (which follows from $f \in L_p[-\pi, \pi]$, some $p > 1$). Let f_n be the trigonometric series estimate of f with parameter m, and let*

$$\lim_{n\to\infty} m = \infty, \qquad \lim_{n\to\infty} \frac{m}{n} = 0.$$

Then $E(\int |f_n - f|) \to 0$ as $n \to \infty$. If we also require that $f \in L_p[-\pi, \pi]$ for some $p > 1$, then $E((f_n - f)^2) \to 0$ as $n \to \infty$ for almost all x.

Proof of Lemma 8 and Theorem 2. The first inequality of Lemma 8 follows from Lemma 1 and a little extra work. The term

$$\frac{1}{\sqrt{n}} \int \sqrt{E\left(D_m^2(x - X_1)\right)}\, dx$$

equals

$$\frac{1}{\sqrt{n}} \sqrt{\frac{2m+1}{2\pi}} \int \sqrt{E\left(F_{2m}(x - X_1)\right)}\, dx$$

$$\leq \sqrt{\frac{2m+1}{2\pi n}}\, \sqrt{2\pi} \sqrt{\int E\left(F_{2m}(x - X_1)\right) dx}$$

$$= \sqrt{\frac{2m+1}{n}},$$

where we used the Cauchy–Schwarz inequality and Lemma 4. In fact, if we use the notation $\sigma_{2m}(f) = E(F_{2m}(x - X_1))$ from Lemma 5, and if we recall

that $\sigma_{2m}(f)$ is a density in x for each m, and $\int|\sigma_{2m}(f)-f|\to 0$, then the same term can be rewritten as

$$\sqrt{\frac{2m+1}{2\pi n}}\int\sqrt{\sigma_{2m}(f)},$$

and

$$\begin{aligned}\int\sqrt{\sigma_{2m}(f)} &= \int\sqrt{(\sigma_{2m}(f)-f)+f}\\ &\le \int\sqrt{|\sigma_{2m}(f)-f|}+\int\sqrt{f}\\ &\le \sqrt{2\pi\int|\sigma_{2m}(f)-f|}+\int\sqrt{f}\\ &= o(1)+\int\sqrt{f},\end{aligned}$$

and this concludes the proof of Lemma 8.

The first part of Theorem 2 follows from Lemmas 7 and 8. This leaves us with the statement about pointwise convergence almost everywhere. By the Carleson–Hunt theorem (Lemma 6, part B), it suffices to show that $E((f_n - E(f_n))^2)\to 0$ for almost all x. But $f_n - E(f_n) = (1/n)\Sigma_{j=1}^n Y_j$, where $Y_j = D_m(x-X_j) - E(D_m(x-X_j))$. It suffices thus that $E(Y_1^2)/n\to 0$ almost everywhere. This is implied in turn by $E(D_m^2(x-X_1))/n\to 0$ almost everywhere. But $D_m^2(u) = ((2m+1)/2)\,F_{2m}(u)$, all u. Thus, it suffices that $m/n\to 0$ and that $E(F_{2m}(x-X_1)) = \sigma_{2m}(f)$ remains bounded for almost all x. This is of course a consequence of the Fejér–Lebesgue theorem (Lemma 5).

Let us finally briefly consider the necessity of the conditions on m. It is clear that if f does not have a finite Fourier series expansion, then $m\to\infty$ is necessary for $\int|S_m(f)-f|\to 0$ and thus for $E(\int|f_n-f|)\to 0$. That there are indeed many densities with a finite Fourier series expansion follows from this simple construction: by using $\cos^2(x) = \frac{1}{2}(1+\cos(2x))$ repeatedly and applying the binomial theorem, it is clear that functions of the form $\cos^{2^r}(x)$, r integer, have an expansion with as highest nonzero Fourier coefficient a_{2^r+1}. Thus, if we normalize these functions to make

them densities, and set m fixed but at least equal to 2^r, then

$$E\left(\int|f_n - f|\right) \le \sqrt{\frac{2m+1}{n}},$$

which decreases to 0 at the rate $1/\sqrt{n}$. It is noteworthy that the kernel estimate cannot achieve this rate for these densities. Of course, this follows from the fact that the trigonometric series estimate is all but tailored for densities of this form.

LEMMA 9 (Achievability of an Error that is $O(1/\sqrt{n})$). *Let f be a density on $[-\pi,\pi]$, and let f_n be the trigonometric series estimate with parameter m. Then* $\limsup_{n\to\infty} E(\int|f_n - f|)\sqrt{n} < \infty$ *implies that* $\limsup_{n\to\infty} m < \infty$.

Proof. It suffices to show that for all densities f on $[-\pi,\pi]$ with $\int|S_m(f) - f| \to 0$ as $m \to \infty$, $\lim_{n\to\infty} m = \infty$ implies

$$E\left(\int|f_n - f|\right) \ge \frac{\log(m)}{\sqrt{n}}(A + o(1))$$

for some universal constant $A > 0$.

We have for all f, as $m \to \infty$,

$$E\left(\int|f_n - f|\right) \ge \frac{1}{2}E\left(\int|f_n - E(f_n)|\right) \quad \text{(Lemma 1)}$$

$$\ge (32n)^{-1/2}\int E\left(|D_m(x - X_1) - E(D_m(x - X_1)|\right)dx$$

(Lemma 5.27)

$$\ge (32n)^{-1/2}\int\left(E(|D_m(x - X_1)|) - |S_m(f)|\right)dx$$

$$\ge (32n)^{-1/2}\left(\int|D_m| - \int|S_m(f) - f| - 1\right).$$

If $\int|S_m(f) - f| \to 0$ as $m \to \infty$, then the lower bound $\sim (4/\pi^2)(\log m)/(\sqrt{32n})$ (Lemma 4). This concludes the proof of Lemma 9.

The necessity of the condition $m/n = o(1)$ for consistency can be obtained with some work from lemma 5.27 as well. This will not be done

here. We should mention here that Bosq and Bleuez (1978) and Bleuez and Bosq (1979) have shown that for densities $f \in L_2[-\pi, \pi]$ having an everywhere convergent Fourier series expansion, the following conditions are equivalent, assuming that $\lim_{n\to\infty} m = \infty$:

(i) $f_n \to f$ in probability, all x, all given f.
(ii) $E(|f_n - f|) \to 0$, all x, all given f.
(iii) $E((f_n - f)^2) \to 0$, all x, all given f.
(iv) $E(\int (f_n - f)^2) \to 0$, all given f.
(v) $\lim_{n\to\infty} m/n = 0$.

With the information available (see, e.g., Lemma 2), the reader should have little difficulty with the proof of this result. In the cited work of Bleuez and Bosq, results of this type are obtained for many orthogonal series estimates as a special case of a very general theorem.

5. THE TRIGONOMETRIC SERIES ESTIMATE: RATE OF CONVERGENCE

From the previous section we can directly conclude that the rates of convergence for the trigonometric series estimate and the kernel estimate are not comparable. For the uniform density on $[-\pi, \pi]$, the trigonometric series estimate has $f_n = f$ when $m = 0$, while no kernel estimate can converge faster than $n^{-1/3}$. The same lower bound applies to densities that are mixtures of the uniform density and densities proportional to $\cos^{2^r}(x)$, r a positive integer, $|x| \le \pi$. For these densities, the trigonometric series estimate achieves the rate $1/\sqrt{n}$ when m remains bounded, $m \ge 2^r$. On the other hand, the trigonometric series estimate is often not even consistent (Theorem 1).

Our first objective here is to show that the trigonometric series estimate has a uniformly bounded L_1 error over the Lipschitz classes $W(s, \alpha, C)$ (see Section 4.2) when m is appropriately chosen, and that the bound comes to within a constant of the minimax lower bound of Theorem 4.6. This property is thus shared with the kernel estimate. We will conclude this section with some remarks about the behavior of the trigonometric series estimate on Bretagnolle–Huber classes and Sobolev classes. In the definition of $W(s, \alpha, C)$, the interval $[0, 1]$ should be replaced by $[-\pi, \pi]$, and the Lipschitz and smoothness conditions imposed on f are assumed to hold on the real line (as in Chapter 4). This is very important, because it forces f to be 0 at $-\pi$ and $+\pi$, and to be sufficiently smooth near the endpoints. We will thus not be interested in Gibbs' phenomenon (see, e.g., Hall, 1981).

The study of rates of convergence for individual f is much more difficult than in the case of the kernel estimate since it depends upon the speed with which $\int|S_m(f) - f|$ tends to 0 (see Lemma 1), and this is not related to standard quantities such as $\int|f^{(r)}|$, at least not directly. Consider for example densities with a finite Fourier series expansion, and one notices that the trouble starts when one wants to obtain lower bounds for the expected L_1 error.

LEMMA 10. *If f is a density in $L_2[-\pi, \pi]$, with Fourier coefficients a_i, then*

$$\int|S_m(f) - f| \le \sqrt{2\pi}\sqrt{\sum_{i=2m+1}^{\infty} a_i^2}.$$

Proof. By the Cauchy–Schwarz inequality and orthonormality,

$$\int|S_m(f) - f| \le \sqrt{2\pi}\sqrt{\int(S_m(f) - f)^2} = \sqrt{2\pi}\sqrt{\sum_{i=2m+1}^{\infty} a_i^2}.$$

LEMMA 11 (Lorentz's Inequality. See also Bary (1964, Vol. 1, pp. 215–217)). *Let $f \in W(0, \alpha, C)$ for some $\alpha \in (0, 1]$, and let its Fourier coefficients be $a_i, i \ge 0$. Then*

$$\sum_{i=2m-1}^{\infty} a_i^2 \le \frac{\gamma C^2}{m^{2\alpha}}, \quad \text{where } \gamma = \frac{\pi^{2\alpha+1}}{(4^\alpha - 1)}, \qquad m \ge 1.$$

Proof. For convenience in notation, we assume that f is periodic with period 2π (define it outside $[-\pi, \pi]$ by periodicity). If f has Fourier series expansion

$$\frac{a_0}{\sqrt{2\pi}} + \sum_{i=1}^{\infty}\left(a_{2i-1}\frac{\cos(ix)}{\sqrt{\pi}} + a_{2i}\frac{\sin(ix)}{\sqrt{\pi}}\right),$$

then $f(x + h) - f(x - h)$ considered as a function of x (h is a constant) has Fourier series expansion

$$2\sum_{i=1}^{\infty}\left(a_{2i}\frac{\sin(ix)}{\sqrt{\pi}} - a_{2i-1}\frac{\cos(ix)}{\sqrt{\pi}}\right)\sin(nh).$$

By Bessel's equality and the Lipschitz condition,

$$4\sum_{i=1}^{\infty}\left(a_{2i-1}^2 + a_{2i}^2\right)\sin^2(ih) = \int_{-\pi}^{\pi}\left(f(x+h) - f(x-h)\right)^2 dx$$

$$\leq \int_{-\pi}^{\pi}\left(C(2h)^{\alpha}\right)^2 dx = 2\pi C^2(2h)^{2\alpha}.$$

Therefore, for any m,

$$\sum_{i=m}^{2m-1}\left(a_{2i-1}^2 + a_{2i}^2\right)\sin^2(ih) \leq \frac{\pi}{2}4^{\alpha}C^2h^{2\alpha}.$$

If we take $h = \pi/4m$ and note that for $m \leq i \leq 2m-1, \sin(ih) \geq \sin(\pi/4) = 1/\sqrt{2}$, we obtain

$$\sum_{i=m}^{2m-1}\left(a_{2i-1}^2 + a_{2i}^2\right) \leq \frac{\pi^{2\alpha+1}C^2}{4^{\alpha}m^{2\alpha}}.$$

Thus,

$$\sum_{i=2m-1}^{\infty} a_i^2 = \sum_{i=m}^{\infty}\left(a_{2i-1}^2 + a_{2i}^2\right)$$

$$= \sum_{j=0}^{\infty}\sum_{i=m2^j}^{m2^{j+1}-1}\left(a_{2i-1}^2 + a_{2i}^2\right)$$

$$\leq \sum_{j=0}^{\infty}\frac{\pi^{2\alpha+1}C^2}{4^{\alpha}\left(m2^j\right)^{2\alpha}}$$

$$= \frac{\pi^{2\alpha+1}C^2}{4^{\alpha}m^{2\alpha}} \cdot \frac{1}{1-4^{-\alpha}}$$

which was to be shown.

LEMMA 12. *Let $f \in W(s,\alpha,C)$ for some integer $s \geq 0$, and some $\alpha \in (0,1]$. In the notation of Lemma* 11,

$$\sum_{i=2m-1}^{\infty} a_i^2 \leq \frac{\gamma C^2}{m^{2s+2\alpha}},$$

where

$$\gamma = \frac{\pi^{2\alpha+1}}{4^{\alpha} - 4^{-s}}, \qquad m \geq 1.$$

Proof. We note first that if f has $s-1$ absolutely continuous derivatives and $f^{(s)}$ is Lipschitz, then differentiation of the Fourier series expansion is formally allowed, and $f^{(s)}$ has a Fourier series expansion

$$\sum_{i=1}^{\infty} i^s \left((-1)^{s/2} a_{2i-1} \frac{\cos(ix)}{\sqrt{\pi}} + (-1)^{s/2} a_{2i} \frac{\sin(ix)}{\sqrt{\pi}} \right), \quad s \text{ even},$$

and

$$\sum_{i=1}^{\infty} i^s \left((-1)^{(s+1)/2} a_{2i-1} \frac{\sin(ix)}{\sqrt{\pi}} + (-1)^{(s-1)/2} \frac{\cos(ix)}{\sqrt{\pi}} \right), \quad s \text{ odd}.$$

By Bessel's equality,

$$\sum_{i=1}^{\infty} (a_{2i-1}^2 + a_{2i}^2) i^{2s} = \int (f^{(s)})^2,$$

an equality that will be useful elsewhere. For s even, one can easily verify that $f^{(s)}(x+h) - f^{(s)}(x-h)$, for h fixed, has Fourier series expansion

$$2 \sum_{i=1}^{\infty} i^s \left((-1)^{s/2} a_{2i-1} \left(\frac{-\sin(ix)}{\sqrt{\pi}} \right) + (-1)^{s/2} a_{2i} \frac{\cos(ix)}{\sqrt{\pi}} \right) \sin(ih).$$

For s odd, we have a similar expression. Thus, arguing as in the proof of Lemma 11,

$$4 \sum_{i=1}^{\infty} (a_{2i-1}^2 + a_{2i}^2) i^{2s} \sin^2(ih) = \int_{-\pi}^{\pi} \left(f^{(s)}(x+h) - f^{(s)}(x-h) \right)^2 dx$$

$$\leq \int_{-\pi}^{\pi} \left(C(2h)^{\alpha} \right)^2 dx = 2\pi C^2 (2h)^{2\alpha}.$$

Thus, as in Lemma 11, we obtain

$$\sum_{i=m}^{2m-1} (a_{2i-1}^2 + a_{2i}^2) i^{2s} \leq \frac{\pi^{2\alpha+1} C^2}{4^{\alpha} m^{2\alpha}}.$$

Therefore,

$$\sum_{i=2m-1}^{\infty} a_i^2 = \sum_{i=m}^{\infty} \left(a_{2i-1}^2 + a_{2i}^2\right)$$

$$\leq \sum_{j=0}^{\infty} \sum_{i=m2^j}^{m2^{j+1}-1} \left(a_{2i-1}^2 + a_{2i}^2\right) i^{2s} \cdot (m2^j)^{-2s}$$

$$\leq \sum_{j=0}^{\infty} \frac{\pi^{2\alpha+1} C^2}{4^\alpha} \cdot (m2^j)^{-2\alpha-2s}$$

$$= \frac{\pi^{2\alpha+1} C^2}{4^\alpha m^{2\alpha+2s}} \frac{1}{1 - 4^{-\alpha-s}},$$

which was to be shown.

THEOREM 3. *Let $\alpha \in (0,1], C > 0$ and $s \geq 0$, s integer, be fixed. Then, for the trigonometric series estimate f_n with parameter m,*

$$\sup_{f \in W(s,\alpha,C)} E\left(\int |f_n - f|\right) \leq \sqrt{\frac{2m+1}{n}} + C\sqrt{2\pi\gamma}\left(\frac{1}{m+1}\right)^{s+\alpha},$$

where

$$\gamma = \frac{\pi^{2\alpha+1}}{4^\alpha - 4^{-s}}$$

is the constant of Lemma 12. In particular, if $m \sim (C\sqrt{\pi\gamma}\, 2(\alpha + s))^{2/(1+2(\alpha+s))} n^{1/(1+2(\alpha+s))}$, then

$$\limsup_{n\to\infty} \sup_{f \in W(s,\alpha,C)} E\left(\int |f_n - f|\right) n^{(\alpha+s)/(1+2(\alpha+s))}$$

$$\leq \left(C\sqrt{\pi\gamma}\, 2(\alpha+s)\right)^{1/(1+2(\alpha+s))} \left(\frac{1+2(\alpha+s)}{2(\alpha+s)}\right)\sqrt{2}.$$

Proof. The first inequality follows from Lemmas 8, 10, and 12. The bound is of the form $u\sqrt{m} + vm^{-(\alpha+s)}$ plus smaller order terms if $\lim_{n\to\infty} m = \infty$.

Here $u = \sqrt{2/n}$, $v = C\sqrt{2\pi\gamma}$. It is minimal (in m) if m is the solution of the equation

$$\frac{u}{2\sqrt{m}} - \frac{v(\alpha + s)}{m^{\alpha+s+1}} = 0.$$

This suggests values of m that are $\sim (2v(\alpha + s)/u)^{2/(1+2(\alpha+s))}$. Resubstitution in the bound gives us our result.

REMARK. For the important class $W(0, 1, C)$, we obtain

$$\limsup_{n\to\infty} \sup_{f\in W(0,1,C)} n^{1/3} E\left(\int |f_n - f|\right) \le \left(2C\frac{\pi^2}{\sqrt{3}}\right)^{1/3}\sqrt{2}$$

when $m \sim (4C^2\pi^4/3n)^{1/3}$. This value should be compared with the lower bound of Theorem 4.7, after replacing the C there by $2\pi C$ (because Theorem 3 is valid for densities on $[-\pi, \pi]$, not $[0, 1]$). In general, we note that the upper bound of Theorem 3 has the same dependence on C and n as the minimax lower bound of Theorem 4.6. To achieve a similar result for the kernel estimate, we had to change the kernel according to s and α. Here, in contrast, only the smoothing parameter m needs adjusting.

From Theorem 3, we conclude that the trigonometric series estimate has the power to achieve any rate of convergence up to $n^{-1/2}$ depending upon the smoothness of f. Bretagnolle–Huber classes are defined in terms of $\int |f^{(s)}|$ and $\int \sqrt{f}$. To achieve a rate of convergence that is asymptotically comparable to that of the minimax lower bound for these classes, we see from Lemma 8 that it suffices to bound $\int |S_m(f) - f|$ from above by an expression that is proportional to $\int |f^{(s)}|/m^s$ under some smoothness conditions on f. Unfortunately, we cannot quite achieve this because we have an extra multiplicative factor of $\log(m)$. To obtain such an upper bound quickly, we can proceed as follows. Let AC_s be the class of all densities on $[-\pi, \pi]$ with $s - 1$ absolutely continuous derivatives (on the real line) and $\int |f^{(s)}| < \infty$. Let T_m be the space of all trigonometric polynomials of degree m, that is, linear functions of $\cos(ix)$ and $\sin(ix)$, $0 \le i \le m$. The, by *Jackson's second theorem* (see, e.g., Butzer and Nessel, 1971, pp. 97–99),

$$\inf_{t_m\in T_m} \int |f - t_m| \le \left(\frac{36}{m}\right)^s \int |f^{(s)}|, \qquad f \in AC_s.$$

This gives us a quick upper bound for the bias:

$$\int |S_m(f) - f| \le \int |S_m(f) - t_m| + \int |t_m - f| \quad (t_m \in T_m)$$

$$\le \int |D_m| \int |f - t_m| + \int |f - t_m| \quad \text{(Young's inequality)}$$

$$\le \left(3 + \frac{4}{\pi^2}\log(m)\right) \int |f - t_m| \quad \text{(Lemma 4)},$$

and by taking the best t_m in T_m, we can conclude the following:

THEOREM 4. *Let f be a density on $[-\pi, \pi]$ in the class AC_s, where $s \ge 0$ is a fixed integer. Then, for the trigonometric series estimate f_n with parameter m,*

$$E\left(\int |f_n - f|\right) \le \sqrt{\frac{2m+1}{n}} + \left(3 + \frac{4}{\pi^2}\log(m)\right)\left(\frac{36}{m}\right)^s \int |f^{(s)}|.$$

The first term on the right-hand side can be replaced by $(\sqrt{m/\pi n})(\int \sqrt{f} + o(1))$.

The question thus arises of whether the extra $\log m$ factor is really necessary. For $s = 0$, we know it is (see, e.g., Theorem 1). For $s > 0$, it seems likely that it cannot be replaced by a smaller factor (see, e.g., Butzer and Nessel (1971, p. 108) or Quade (1937)). It is precisely this obstacle that has kept several researchers from studying the performance of the trigonometric series estimate in terms of $\int |f^{(s)}|$. Wahba (1975) for example considers *Sobolev spaces*, that is, spaces of densities f with $s - 1$ absolutely continuous derivatives, and $\int |f^{(s)}|^p \le M < \infty$, where $p > 1$ is another parameter defining the Sobolev space. In her famous study, she compares the performances of several density estimates in these spaces. For the trigonometric series estimate, all cases $p > 1$ can be handled easily via the Hausdorff–Young inequality (see Bary, 1964, Vol. 1, p. 218) which links the qth norm of the Fourier coefficients with the pth norm of a function, where $1/p + 1/p = 1$. For $p = 2$, this reduces to Bessel's equality, and it is worthwhile to show for this special case how one can proceed.

LEMMA 13. *Let f be an absolutely continuous density with support contained in $[0, 1]$, and let $\int f'^2 < \infty$. Then, for the trigonometric series estimate*

f_n with parameter m,

$$E\left(\int |f_n - f|\right) \le \sqrt{\frac{2m+1}{n}} + \frac{\gamma\sqrt{\int f'^2}}{m+1},$$

where $\gamma = \pi^{3/2}/\sqrt{3}$.

REMARK. The upper bound of Lemma 13 decreases at the rate $O(n^{-1/3})$ when m increases as $n^{1/3}$. By generalizing the argument given below in the manner of the extension of Lemma 11 in Lemma 12, we can treat all Sobolev spaces with $p = 2$ (and obtain bounds in terms of $\int (f^{(s)})^2$). A quick comparison with Theorem 4 shows that we have effectively eliminated the $\log m$ factor at the expense of an additional condition, that is, $f^{(s)} \in L_2[-\pi, \pi]$.

Proof of Lemma 13. We will use Lemmas 8 and 10. In addition, a replacement is needed for Lemma 11. In the proof of Lemma 11, the expression that needs to be treated differently is

$$\int_{-\pi}^{\pi} (f(x+h) - f(x-h))^2 \, dx.$$

Again making f periodic, we see that this equals

$$\int_{-\pi}^{\pi} \left(\int_{x-h}^{x+h} f'\right)^2 dx \le \int_{-\pi}^{\pi} 2h \left(\int_{x-h}^{x+h} f'^2\right) dx$$

$$= (2h)^2 \int f'^2.$$

Thus, the remainder of Lemma 11 can be repeated if we formally replace α by 1 and $2\pi C^2$ by $\int f'^2$. In particular,

$$\sum_{i=2m-1}^{\infty} a_i^2 \le \frac{\pi^3}{3} \frac{\int f'^2}{2\pi} \frac{1}{m^2} = \frac{\pi^2}{6} \frac{\int f'^2}{m^2}.$$

This concludes the proof of Lemma 13.

6. THE HERMITE SERIES ESTIMATE

On the real line, the Hermite series estimate is without any doubt the most popular orthogonal series estimate (Schwartz, 1967; Bosq and Bleuez, 1978; Bleuez and Bosq, 1979; Walter, 1977; Greblicki, 1981). In this section, we will briefly analyze its main properties. Nearly everything that is said below remains valid for the Laguerre series estimate on $[0, \infty)$.

THEOREM 5 (Nonconsistency of the Hermite Series Estimate). *The Hermite orthonormal system is not a basis for L_p for any $p \in [1, \frac{4}{3}]$, or $p \in [4, \infty)$. If f_n is the Hermite series estimate with parameter m_n and partial sum $S_{m_n}(f)$, then, provided that*

$$\lim_{n \to \infty} m_n = \infty, |m_n - m_{n-1}| \leq 1, \quad \text{all } n,$$

we can find a density f such that

$$\limsup_{n \to \infty} \int |S_{m_n}(f) - f| = \infty,$$

and thus

$$\limsup_{n \to \infty} E\left(\int |f_n - f| \right) = \infty.$$

That the condition on L_p cannot be removed easily is seen from the following classical result in analysis:

LEMMA 14 (Askey and Wainger, 1965; see also, Muckenhoupt, 1970). *For all $f \in L_p$, and all $p \in (\frac{4}{3}, 4)$,*

$$\lim_{m \to \infty} \int |S_m(f) - f|^p = 0.$$

For all $p \notin (\frac{4}{3}, 4)$, there exists an $f \in L_p$ such that

$$\limsup_{m \to \infty} \int |S_m(f) - f|^p > 0.$$

LEMMA 15 (Skovgaard's Bounds; see Askey and Wainger, 1965, p. 700). *Let p_i be the ith function in the Hermite orthonormal system. Then there exist*

positive constants C_1, C_2, C_3, C_4 not depending upon i or x such that

$$|p_i(x)| \le \begin{cases} C_1 i^{-1/12}, & \text{all } x, i, \\ C_2 i^{-1/4}, & \text{all } |x| \le \sqrt{4i}\,, \big||x| - \sqrt{2i}\big| \ge (2i)^{-1/6}, \\ C_2 \exp(-C_4 x^2), & \text{all } |x| \ge \sqrt{4i}\,. \end{cases}$$

Proof of Theorem 5. Because S_m is a linear operator, we have for any functions f_1, f_2,

$$\int |S_m(f_1 + f_2) - (f_1 + f_2)| \le \int |S_m(f_1) - f_1| + \int |S_m(f_2) - f_2|.$$

Taking $f_1 = (f)_+$, $f_2 = (f)_-$ for some $f \in L_1$, it is easily seen that it suffices to prove that $\limsup_{n\to\infty} \int |S_{m_n}(f) - f| = \infty$ for some $f \in L_1$. By the uniform boundedness principle (see, e.g., Butzer and Nessel, 1971, pp. 18–19), we are done if we can show that

$$\sup_m \sup_{f \in L_1} \frac{\int |S_m(f)|}{\int |f|} = \infty.$$

We will argue now by contradiction. If we assume that there exists a finite M such that $\int |S_m(f)| \le M \int |f|$, all $m, f \in L_1$, then

$$\int |S_m(f) - S_{m-1}(f)| = \int |a_m p_m| \le 2M \int |f|.$$

But $\int |a_m p_m| = |\int f p_m| \int |p_m|$. We know that for some $f \in L_1$,

$$\left| \int f p_m \right| \ge \frac{1}{2} \int |f| \operatorname{ess\,sup} |p_m|$$

and thus we would have

$$\operatorname{ess\,sup} |p_m| \int |p_m| \le 4M.$$

This is clearly impossible in view of $\operatorname{ess\,sup} |p_m| \ge c m^{-1/12}$ all m large enough, some $c > 0$ (the supremum being reached near $\sqrt{2m}$), and $\int |p_m| \ge c m^{-1/4}$, $m \ge 1$, for another constant $c > 0$. (These relations do not follow

from Skovgaard's bounds, but they would if the bounds were sharp, and they are (Askey and Wainger, 1965).) This gives us our contradiction. If we replace m by m_n and require that m_n diverges and $|m_n - m_{n-1}|$ does not exceed 1 for any n, then $\limsup_{n\to\infty} \int |S_{m_n}(f) - f| = \infty$ for some $f \in L_1$.

The first statement of Theorem 5 will not be proved here.

LEMMA 16. *Let f_n be the Hermite series estimate based on $X_1, \ldots, X_n$, an independent sample drawn from density f. Then, for all $n \geq 1$, and all parameters $m \geq 1$,*

$$\lim_{a\to\infty} \int |f_n(x + a, X_1 + a, \ldots, X_n + a)| = 0 \quad \textit{almost surely},$$

and

$$\lim_{a\to\infty} E\left(\int |f_n(x + a, X_1 + a, \ldots, X_n + a)| \right) = 0.$$

Proof.

$$\int |f_n(x + a, X_1 + a, \ldots, X_n + a)| \leq \sum_{i=0}^{m} \left| \frac{1}{n} \sum_{j=1}^{n} p_i(X_j + a) \right| \int |p_i(x + a)|$$

$$\leq \frac{1}{n} \sum_{j=1}^{n} \sum_{i=0}^{m} |p_i(X_j + a)| \int |p_i|$$

$$\leq (m + 1) \sup_i \int |p_i| \sup_{i,j} |p_i(X_j + a)|.$$

But $\sup_i \int |p_i|$ is finite by Skovgaard's bounds (Lemma 15), and $\sup_{i,j} |p_i(X_j + a)| \to 0$ as $a \to \infty$, again by Lemma 15. This concludes the proof of Lemma 16.

That the Hermite series estimate could not possibly be translation invariant follows from Lemma 3 about orthogonal series estimates on the real line. The property given in Lemma 16 is puzzling since we have no guarantees whatsoever as to the size of $\int |f_n|$, let alone $\int f_n$. Together with Theorem 5, we must conclude that the Hermite series estimate seems ill suited as a general purpose density estimate on the real line.

The Hermite series estimate is consistent in L_p for all $p \in (\frac{4}{3}, 4)$, $f \in L_p$. In Lemma 17, we will show this for $p = 2$. The problem with obtaining convergence in L_1 is that we cannot obtain an upper bound for $\int |S_m(f) - f|$

from some L_p norm of $S_m(f) - f$ by Hölder's inequality because we are integrating over R. We can come close to L_1 convergence via other devices. For example, Muckenhoupt (1970) has shown that for fixed $b > 0$, $B \geq \max(b + \frac{1}{3}, -\frac{2}{3})$,

$$\int \frac{|S_m(f) - f|}{|1 + |x||^b} \to 0 \quad \text{as } m \to \infty,$$

provided that $\int f \log_+ f < \infty$, $\int f|x|^B \log_+ f < \infty$, $\int f|x|^{B+2} < \infty$. Unfortunately, in his result, we cannot set $b = 0$, $B = \frac{1}{3}$.

LEMMA 17 (L_2 Convergence of the Hermite Series Estimate). *Let f_n be the Hermite series estimate with parameter m, and let*

$$\lim_{n \to \infty} m = \infty; \quad \lim_{n \to \infty} \frac{m}{n^2} = 0.$$

Then, for all densities f in L_2, $E(\int (f_n - f)^2) \to 0$ as $n \to \infty$.

Proof. By Lemma 2 and Bessel's equality (see also Lemma 14), we see that it suffices to show that the variance term $E(\int (f_n - E(f_n))^2)$ tends to 0. This follows from

$$\int f \sum_{i=0}^{m} p_i^2 = o(n).$$

By Skovgaard's bounds, for $|x| \geq \sqrt{4m}$,

$$\sum_{i=0}^{m} p_i^2 \leq (m+1)C_3^2 \exp(-2C_4 x^2) \leq (m+1)C_3^2 \exp(-8C_4 m)$$

so that

$$\int_{|x| \geq \sqrt{4m}} f \sum_{i=0}^{m} p_i^2 = O(e^{-C_4 m}) = o(n).$$

Define next

$$q_m(x) = \sum_{i=1}^{m} \left(C_1 i^{-1/12}\right)^2 I_{[|x| \leq \sqrt{4m},\, ||x| - \sqrt{2i}| \leq (2i)^{-1/6}]}.$$

Then, by Lemma 15,

$$\int_{|x|\le\sqrt{4m}} f\sum_{i=1}^{m} p_i^2 \le \int f q_m + \sum_{i=1}^{m}\left(C_2 i^{-1/4}\right)^2 + \sum_{i=1}^{m} C_3^2 \exp(-8C_4 i)$$

$$= I_1 + I_2 + I_3.$$

Clearly, $I_3 = O(1)$. Also, $I_2 = O(\sqrt{m}) = o(n)$. Finally, with a little work one can show that the indicator function in the definition of $q_m(x)$ is nonzero for almost $C_5 m^{1/3}$ indices, uniformly over all $|x| \le \sqrt{4m}$, where $C_5 > 0$ is a given constant. Thus, uniformly over such x, $q_m(x) = O(m^{5/18}) = o(n)$, and we are done.

REMARK. Lemma 17 was obtained by Schwartz under the stronger condition $m = o(n)$ (Schwartz, 1967), and by Greblicki (1981) under the condition $m = o(n^{6/5})$. The necessity of the conditions on m given in Lemma 17, if convergence is to hold for all $f \in L_2$, was obtained by Bleuez and Bosq (1979).

REMARK. We have not discussed the pointwise convergence of the Hermite series estimate up to this point. Based on the Carleson–Hunt theorem, Muckenhoupt (1970) proved that $S_m(f) \to f$ at almost all x when $\int f (\log_+ f)^2 < \infty$. This, together with Skovgaard's bounds, can be used to obtain the pointwise convergence of this estimate.

REMARK. The Laguerre series estimate behaves very much like the Hermite series estimate. For important references on the bias term, see Askey and Wainger (1965) and Muckenhoupt (1970a, b, c).

7. THE LEGENDRE SERIES ESTIMATE

The Legendre series estimate was suggested for use in density estimation by Crain (1974) and Hall (1982) (see also Viollaz, 1980). In many respects, the estimate converges and diverges under circumstances that are similar to the Hermite series estimate. This is apparent when one compares the following lemma with Lemma 14.

LEMMA 18. *The Legendre polynomials form a basis for* $L_p[-1,1]$ *if* $\frac{4}{3} < p < 4$, *and only for those* p. *For* $f \in L_p[-1,1]$, $\frac{4}{3} < p < 4$,

$$\int |S_m(f) - f|^p \to 0 \quad \text{as } m \to \infty.$$

For all $p \notin (\frac{4}{3}, 4)$, *there exists an* f *in* $L_p[-1, 1]$ *such that*

$$\limsup_{m \to \infty} \int |S_m(f) - f|^p > 0.$$

Lemma 18 is due to Pollard (1947) and Newman and Rudin (1952). For its extension to Jacobi polynomials, see Pollard (1948, 1949). See also Muckenhoupt (1969). Since we are mainly interested in L_1, it is perhaps of interest to exhibit a density that cannot be estimated by the Legendre series estimate.

THEOREM 6 (Nonconsistency of the Legendre Series Estimate). *Let us consider the density*

$$f(x) = \frac{7/4}{2^{7/4}}(1 - x)^{-3/4}, \qquad |x| < 1.$$

Then f *is in* $L_p[-1, 1]$ *for all* $p \in [1, \frac{4}{3})$,

$$\liminf_{m \to \infty} \left(\int |S_m(f) - f| + \int |S_{m+1}(f) - f| \right) > 0,$$

$$\limsup_{m \to \infty} \int |S_m(f) - f| > 0,$$

and

$$\inf_{n,m} \left(E\left(\int |f_{n,m} - f| \right) + E\left(\int |f_{n,m+1} - f| \right) \right) > 0,$$

where $f_{n,m}$ *is the Legendre series estimate with parameter* m *and sample size* n.

Proof. It is easy to verify that $f \in L_p, 1 \le p < 4/3$. The remainder of the Theorem follows from Lemma 1,

$$\int |S_m(f) - f| + \int |S_{m-1}(f) - f| \ge \int |S_m(f) - S_{m-1}(f)| = |a_m| \int |p_m|,$$

$a_m \ge A_1 + o(1)$, $\int |p_m| \ge A_2 + o(1)$, as $m \to \infty$, where A_1, A_2 are positive constants (see Szegö, 1975, p. 256 and p. 173, respectively).

REMARK. In Theorem 6, we stopped short of showing that for the given density

$$\inf_{n,m} E\left(\int |f_{n,m} - f|\right) > 0.$$

This requires a much more sophisticated argument.

THEOREM 7 (Consistency of the Legendre Series Estimate). *If f is a density on $[-1,1]$, $f \in L_p[-1,1]$ for some $p > 4/3$,*

$$\int_{-1}^{1} (1-x^2)^{-1/2} f(x)\,dx < \infty,$$

and

$$\lim_{n\to\infty} m = \infty, \quad \lim_{n\to\infty} \frac{m}{n} = 0,$$

then

$$\lim_{n\to\infty} E\left(\int |f_n - f|\right) = 0,$$

where f_n is the Legendre series estimate with parameter m.

Proof. Define q by $1/p + 1/q = 1$. If $p > 4$, replace p by 2 first. By Lemma 18, and Hölder's inequality,

$$\int |S_m(f) - f| \le 2^{1/q}\left(\int |S_m(f) - f|^p\right)^{1/p} \to 0 \quad \text{as } m \to \infty.$$

By Lemma 1, the Cauchy–Schwarz inequality and Lemma 2, we have

$$E\left(\int |f_n - E(f_n)|\right) \le \int \sqrt{E\left((f_n - E(f_n))^2\right)}$$

$$\le \sqrt{2\int E\left((f_n - E(f_n))^2\right)}$$

$$\le \sqrt{2}\sqrt{\frac{1}{n}\int f \sum_{i=0}^{m} p_i^2}.$$

By Stieltjes' first theorem (Sansone, 1977, p. 199),

$$|p_i| \le \sqrt{\frac{2i+1}{2}}\, 4\sqrt{\frac{2}{\pi}}\, \frac{1}{\sqrt{i}}(1-x^2)^{-1/4}$$

$$\le \sqrt{1+\frac{1}{i}}\, 4\sqrt{\frac{2}{\pi}}\,(1-x^2)^{-1/4}$$

$$\le \frac{8}{\sqrt{\pi}}(1-x^2)^{-1/4}, \qquad i \ge 1.$$

Since $p_0 = 1/\sqrt{2}$, the bound is valid for $i = 0$ too. Thus,

$$\sum_{i=0}^{m} p_i^2 \le (m+1)\frac{64}{\pi}(1-x^2)^{-1/2},$$

so that we can conclude that $E(\int |f_n - E(f_n)|) \to 0$.

None of the conditions of convergence in Theorem 7 can be entirely omitted. Hall (1982) gives convincing arguments in favor of the Legendre series estimate, for example, very few terms are needed to achieve good rates of convergence in $L_2[-1, 1]$ for certain classes of densities. The rate of convergence will not be dealt with here. It is perhaps noteworthy too that the estimate is not translation invariant, but that nevertheless $\int f_n = 1$, all n, m.

8. SINGULAR INTEGRAL ESTIMATES

The *singular integral estimate* of f with *kernel* K_m is defined by

$$f_n(x) = \frac{1}{n}\sum_{j=1}^{n} K_m(x - X_j).$$

Estimates of this form include the kernel estimate, and the trigonometric series estimate (where the kernel is the Dirichlet kernel). We could have written $K_m(x, X_j)$ to gain generality (such estimates are called Dirac delta function estimates under some conditions on K_m, see Walter and Blum (1979)), but translation invariance forces to consider the case $K_m(x - X_j)$

only. We will treat estimation of densities on $[-\pi, \pi]$ only, and impose some conditions on K_m:

$$K_m(x) = K_m(-x); \quad \int_{-\pi}^{\pi} K_m = 1; \quad K_m \text{ is periodic with period } 2\pi;$$

$$\int_{-\pi}^{\pi} |K_m| < \infty. \tag{1}$$

In view of the periodicity of K_m, the integral of K_m over the real line would be ∞, and the kernel estimate is thus no longer a special case of the singular integral estimate with kernels satisfying (1).

Singular integral estimates are given the standard treatment: first we will show that we can choose sequences of kernels such that the singular integral estimate is consistent for all $f \in L_1[-\pi, \pi]$. We can even choose all the K_m's nonnegative, so that f_n is a density on $[-\pi, \pi]$, all n, m. Then, we will analyze the rate of convergence, and observe for example that if $K_m \geq 0$, the expected L_1 error cannot decrease to 0 faster than $n^{-2/5}$, just as for the kernel estimate.

First, a few definitions are in order. We define the *singular integral* $S_m(f)$ (or $S_m(f, x)$) by

$$S_m(f) = \int_{-\pi}^{\pi} f(x - u) K_m(u)\, du,$$

where f is extrapolated over R by periodicity. Thus, Young's inequality remains valid:

$$\int |S_m(f)| \leq \int |f| \int |K_m|.$$

Following Butzer and Nessel (1971, p. 31), we say that K_m is an *approximate identity* when (1), (2), and (3) hold:

$$\sup_m \int |K_m| \leq C < \infty; \tag{2}$$

$$\lim_{m \to \infty} \int_{\delta \leq |u| \leq \pi} |K_m(u)|\, du = 0, \quad \text{all } \delta > 0. \tag{3}$$

K_m is a *strong approximate identity* if (3) is replaced by

$$\lim_{m\to\infty} \sup_{\delta\le|u|\le\pi} |K_m(u)| = 0, \quad \text{all } \delta > 0. \tag{4}$$

LEMMA 19. *For all approximate identities and all* $f \in L_1[-\pi,\pi]$,

$$\int |S_m(f) - f| \to 0 \quad \text{as } m \to \infty.$$

Proof.

$$S_m(f) - f = \int_{-\pi}^{\pi} (f(x-u) - f(x)) K_m(u)\, du.$$

Let $g(u)$ be $\int_{-\pi}^{\pi} |f(x-u) - f(x)|\, dx$. By Young's inequality,

$$\begin{aligned}\int |S_m(f) - f| &\le \int_{-\pi}^{\pi} g(u)|K_m(u)|\, du \\ &\le \int_{|u|\le\delta} \sup_{v\le\delta} g(v)|K_m(u)|\, du \\ &\quad + \int_{\delta\le|u|\le\pi} |K_m(u)|\, du \cdot 2\int|f| \\ &\le C \sup_{v\le\delta} g(v) + o(1).\end{aligned}$$

But we know that $\lim_{v\downarrow 0} g(v) = 0$ for all $f \in L_1[-\pi,\pi]$. This concludes the proof of Lemma 19.

THEOREM 8 (Consistency of Singular Integral Estimates). *Let* f_n *be a singular integral estimate with parameter* m, *where* $\lim_{n\to\infty} m = \infty$. *The sequence of kernels* K_m *is assumed to be an approximate identity, and* $\int K_m^2 = o(n)$ *as* $n \to \infty$. *Then*

$$E\left(\int |f_n - f|\right) \to 0 \quad \text{as } n \to \infty, \quad \text{all densities } f \text{ on } [-\pi,\pi].$$

Proof. By Lemma 19, it suffices to prove that $E(\int |f_n - E(f_n)|) \to 0$ as $n \to \infty$. (Note that $E(f_n) = S_m(f)$.) By a double application of the

Cauchy–Schwarz inequality, we have

$$E\left(\int |f_n - E(f_n)|\right) = \int E\left(\left|\frac{1}{n}\sum_{j=1}^{n}\left(K_m(x - X_j) - E\left(K_m(x - X_j)\right)\right)\right|\right)$$

$$\leq \int \sqrt{E\left(\frac{1}{n}\mathrm{Var}\left(K_m(x - X_1)\right)\right)}$$

$$\leq \sqrt{2\pi}\sqrt{\frac{1}{n}\int \mathrm{Var}\left(K_m(x - X_1)\right)}$$

$$\leq \sqrt{\frac{2\pi}{n}}\sqrt{\int E\left(K_m^2(x - X_1)\right)}$$

$$= \sqrt{\frac{2\pi}{n}}\sqrt{\int K_m^2},$$

where all integrals are over $[-\pi, \pi]$. This concludes the proof of Theorem 8.

We have already encountered an approximate identity. For example, the *Fejér kernel*

$$K_m(u) = F_m(u) = \frac{1}{2\pi(m+1)}\left(\frac{\sin((m+1)u/2)}{\sin(u/2)}\right)^2$$

is nonnegative, and defines an approximate identity. To see this, we recall from Lemma 4 that $\int F_m = 1$, $\int |F_m| = 1$, and that $\int_{\delta \leq |u| \leq \pi} F_m(u)\,du \leq \int_{\delta \leq |u| \leq \pi} \pi(2mu^2)^{-1}\,du \leq \pi(2m\delta)^{-1} \to 0$ as $m \to \infty$, all $\delta > 0$.

On the contrary, D_m is not an approximate identity in view of Theorem 1 and/or Lemma 4. Here is a short list of several approximate identities—all but two of these are also nonnegative:

(i) The *Rogosinski kernel* (Butzer and Nessel, 1971, pp. 56ff)

$$\frac{1}{2}\left(D_m\left(x + \frac{\pi}{2m+1}\right) + D_m\left(x - \frac{\pi}{2m+1}\right)\right).$$

(ii) *Jackson's kernel* (Butzer and Nessel, 1971, pp. 60–61)

$$\frac{3}{2\pi m(2m^2+1)}\left(\frac{\sin(mx/2)}{\sin(x/2)}\right)^4.$$

(iii) The *Fejér–Korovkin kernel* (Butzer and Nessel, 1971, pp. 79–80)

$$\frac{\sin^2(\pi/(m+2))}{\pi(m+2)}\left(\frac{\cos((m+2)x/2)}{\cos(\pi/(m+2))-\cos(x)}\right)^2.$$

(iv) The *de la Vallée Poussin kernel* (Butzer and Nessel, 1971, p. 112)

$$\frac{m!^2}{(2\pi)(2m)!}\left(2\cos\left(\frac{x}{2}\right)\right)^{2m}.$$

(v) The *Jackson–de la Vallée Poussin kernel* (Butzer and Nessel, 1971, p. 131)

$$\frac{2+\cos(x)}{4\pi m^3}\left(\frac{\sin(mx/2)}{\sin(x/2)}\right)^4.$$

(vi) *de la Vallée Poussin's second kernel* (Butzer and Nessel, 1971, p. 108)

$$(1+2\cos(mu))F_{m-1}(u).$$

The Fejér, Rogosinski, and Fejér–Korovkin singular integral estimates were analyzed and compared (from an L_2 point of view) by Hall (1983). Some properties of the Fejér singular integral estimate were obtained by Krzyzak and Pawlak (1982). Under certain continuity conditions on f, the Rogosinski singular integral estimate achieves for example the same rate of convergence for the expected L_2 error as the kernel estimate, but has a smaller asymptotic constant, provided that m is chosen optimally (Hall, 1983). This intriguing observation adds a little excitement to the study of the singular integral estimates, and will motivate us in our analysis of the rate of convergence in L_1. Before we do so, we would like to point out the close connection between singular integral estimates and the trigonometric series estimate.

The singular integral estimates can often be written as

$$f_n(x) = \frac{1}{2\pi} + \sum_{i=1}^{\infty}\left(a_{n,2i-1}\frac{\cos(ix)}{\sqrt{\pi}} + a_{n,2i}\frac{\sin(ix)}{\sqrt{\pi}}\right)\lambda_{mi}\sqrt{\pi},$$

where the a_{ni}'s are as for the trigonometric series estimate, that is, they are the standard estimates of the Fourier coefficients a_i of f for the trigonomet-

ric system. This form of the singular integral estimates follows from the decomposition

$$K_m(u) = \frac{1}{2\pi} + \sum_{i=1}^{\infty} \lambda_{mi} \frac{\cos(iu)}{\sqrt{\pi}},$$

valid for all even kernels with a convergent Fourier series (use the formula $\cos(i(x-y)) = \cos(ix)\cos(iy) + \sin(ix)\sin(iy)$). The λ_{mi}'s in the definition of f_n are thus the Fourier coefficients of the kernel K_m.

The given form of f_n is of particular interest to the practitioner when the λ_{mi}'s are zero for all i large enough: this would provide the flexibility of computing and storing the a_{ni}'s instead of the X_j's.

Watson (1969) and Rosenblatt (1971) have suggested the smoothed orthogonal series estimators $\sum_{i=0}^{\infty} \lambda_{mi} a_{ni} p_i$, where the weights λ_{mi} play the role of smoothers. For the choice $\lambda_{mi} = 1$, $i \le m$, $\lambda_{mi} = 0$, $i > m$, it is clear that no smoothing is done, and we obtain the orthogonal series estimate again. Smoothing has may uses: roughly speaking, if we smooth well, we will obtain consistency in L_1 for all densities. A case in point is the singular integral estimate with an approximate identity. But we loose in fine-tuning, that is, for certain classes of densities, the rate of convergence of the smoothed estimate is inferior to that of the original orthogonal series estimate. This too will be illustrated in this section. Interestingly, Watson (1969) has obtained the best form for λ_{mi} for fixed n when one wants to minimize $E(\int (f_n - f)^2)$:

$$\lambda_{mi} = \frac{a_i^2}{a_i^2 + (1/n)\left(\int f p_i^2 - a_i^2\right)}$$

(this can be verified in two lines). Unfortunately, this is of little help, since the a_i's are unknown, and because we are dealing with the L_2 error. Various suggestions have been made in the literature as to how the λ_{mi}'s should be chosen both in general and for particular orthonormal systems, see, for example, Whittle (1958), Fellner (1974), Brunk (1977, 1978), Kronmal and Tarter (1968), and Wahba (1978). The motivation for these suggestions is often different. Brunk, for example, uses a Bayesian argument, and allows for the use of a priori information in λ_{mi}'s of the form $c_i/(c_i + 1/n)$, $1 \le i \le n$. Wahba (1978) discusses estimators with $\lambda_{mi} = (1 + ci^p)^{-1}$, $1 \le i \le n$, where $c, p > 0$ are constants. From Watson's formula, we can immediately think of automatic methods for choosing the λ_{mi}'s. The automatization of orthogonal series estimates or their smoothed versions is

dealt with, for example, in Kronmal and Tarter (1968), Tarter and Kronmal (1976), Crain (1973), Asselin de Beauville (1978), and Wahba (1977, 1978).

Let us return now to the singular integral estimates, that is, smoothed trigonometric series estimates. If we consider the Dirichlet kernel

$$D_m(u) = \frac{1}{2\pi} + \sum_{i=1}^{m} \frac{1}{\pi}\cos(ix),$$

we note immediately that

$$\lambda_{mi} = \begin{cases} 1/\sqrt{\pi}, & 1 \le i \le m, \\ 0, & i > m. \end{cases}$$

Here are a few other examples:

(i) *The Fejér kernel*:

$$\lambda_{mi}\sqrt{\pi} = \begin{cases} \left(1 - \frac{i}{m+1}\right), & 1 \le i \le m \\ 0, & i > m. \end{cases}$$

(ii) *Rogosinski's kernel*:

$$\lambda_{mi}\sqrt{\pi} = \begin{cases} \cos\left(\frac{i\pi}{2m+1}\right), & 1 \le i \le m, \\ 0, & i > m. \end{cases}$$

(iii) *The Fejér–Korovkin kernel*:

$$\lambda_{mi}\sqrt{\pi} = \begin{cases} \dfrac{(m-i+3)\sin\left(\pi\frac{i+1}{m+2}\right) - (m-i+1)\sin\left(\pi\frac{i-1}{m+2}\right)}{2(m+2)\sin\left(\frac{\pi}{m+2}\right)}, & 1 \le i \le m, \\ 0, & i > m. \end{cases}$$

(iv) *de la Vallée Poussin's kernel*:

$$\lambda_{mi}\sqrt{\pi} = \begin{cases} \dfrac{m!^2}{(m-i)!(m+i)!}, & 1 \le i \le m, \\ 0, & i > m. \end{cases}$$

(v) *de la Vallée Poussin's second kernel*:

$$\lambda_{mi}\sqrt{\pi} = \begin{cases} 1, & 1 \le i \le m, \\ 2 - \dfrac{i}{2m+1}, & m < i \le 2m-1 \\ 0, & i > 2m. \end{cases}$$

The weights λ_{mi} are of crucial importance in the study of the rate of convergence of singular integral estimates. The fundamental inequality where one can start from is given in Lemma 20:

LEMMA 20. *Let $S_m(f), S_m^*(f)$ be the singular integrals for f with kernels K_m and $K_m^2/\int K_m^2$, respectively, and let f_n be the singular integral estimate with kernel K_m. Then, for all densities f on $[-\pi, \pi]$,*

$$\int |S_m(f) - f| \le E\left(\int |f_n - f|\right)$$

$$\le \int |S_m(f) - f| + \frac{\int \sqrt{f} \sqrt{\int K_m^2}}{\sqrt{n}}$$

$$+ \left(\frac{\int K_m^2}{n} 2\pi \int |S_m^*(f) - f|\right)^{1/2}.$$

The upper bound is $\int |S_m(f) - f| + (\int \sqrt{f} + o(1))\ (\int K_m^2/n)^{1/2}$ whenever $K_m^2/\int K_m^2$ is an approximate identity. All integrals are over $[-\pi, \pi]$.

Proof. Let us return to the proof of Theorem 8, and note the inequality

$$E\left(\int |f_n - E(f_n)|\right) \le \int \sqrt{E\left(\frac{1}{n} K_m^2(x - X_1)\right)}\, dx$$

$$= n^{-1/2} \int \sqrt{\int K_m^2(x-y) f(y)\, dy}\, dx$$

$$\le n^{-1/2} \int \sqrt{f} \sqrt{\int K_m^2}$$

$$+ n^{-1/2} \int \sqrt{\left| \int K_m^2(x-y)(f(y) - f(x))\, dy \right|}\, dx.$$

The last term does not exceed

$$n^{-1/2}\sqrt{\int K_m^2}\int\sqrt{|S_m^*(f)-f|} \le n^{-1/2}\sqrt{\int K_m^2}\sqrt{2\pi}\sqrt{\int|S_m^*(f)-f|},$$

and this is o(first term) if $K_m^2/\int K_m^2$ is an approximate identity (Lemma 19).

REMARK. If K_m is an approximate identity, then $K_m^2/\int K_m^2$ is an approximate identity whenever

$$\sup_m \sup_u \frac{K_m(u)}{\int K_m^2} < \infty. \tag{6}$$

From here onward, we will only be concerned with $\int|S_m(f)-f|$, and will leave the standard argument of choosing m to minimize the upper bound of Lemma 20 to the reader. From Lemma 20, we can obtain of course both uniform and individual bounds. For example, if $\int|S_m(f)-f| = O(m^{-\alpha})$ and $\sup_u|K_m(u)| = O(m^{\beta})$ for some $\alpha, \beta \ge 0$, then the choice $m \sim n^{1/(\beta+2\alpha)}$ gives

$$E\left(\int|f_n - f|\right) = O(n^{-\alpha/(\beta+2\alpha)}).$$

For all the kernels, except de la Vallée Poussin's (when $\beta = \frac{1}{2}$), we have $\beta = 1$. The important values for α to remember are $\alpha = 1$ and $\alpha = 2$. For the kernels with $\beta = 1$, the rates of convergence that are attainable are $n^{-1/3}$ and $n^{-2/5}$ respectively. The remainder of this section is largely devoted to the computation of α for large classes of densities and large classes of kernels.

We recall first that AC_s is the class of all functions $f \in L_1[-\pi,\pi]$ with $s-1$ absolutely continuous derivatives and $\int|f^{(s)}| < \infty$. While everything that will be said below is valid when these conditions hold for the periodically continued version of f, we will assume that they hold on the real line for the original f; this will allow us further on to make meaningful comparisons with other estimates.

Following Butzer and Nessel (1971), we define the L_1 *modulus of continuity* for a function $f \in L_1[-\pi,\pi]$ by

$$\omega(f,\delta) = \sup_{|h|\le\delta}\int_{-\pi}^{\pi}|f(x+h)-f(x)|\,dx, \qquad \delta > 0,$$

and the *second L_1 modulus of continuity* by

$$\omega^*(f,\delta) = \sup_{|h|\le\delta} \int_{-\pi}^{\pi} |f(x+h) + f(x-h) - 2f(x)|\,dx, \qquad \delta > 0.$$

We also recall the definition of the Lipschitz classes $W(s,\alpha,C)$ for $\alpha \in (0,1]$, $C > 0$, $s \ge 0$ integer: $W(s,\alpha,C)$ is the class of all densities in AC_s for which

$$|f^{(s)}(x) - f^{(s)}(y)| \le C|x-y|^\alpha, \quad \text{all } x, y \in R.$$

LEMMA 21 (Bounds for the L_1 Moduli of Continuity).

A. $\omega^*(f,\delta) \le 2\omega(f,\delta)$, *all* $\delta > 0$. *This tends to* 0 *as* $\delta \downarrow 0$, *for all* $f \in L_1[-\pi,\pi]$.

B. *For* $f \in W(0,\alpha,C)$:

$$\omega(f,\delta) \le 2\pi C\delta^\alpha, \omega^*(f,\delta) \le 4\pi C\delta^\alpha, \delta > 0.$$

C. *For* $f \in W(1,\alpha,C)$:

$$\omega^*(f,\delta) \le 2\pi C\delta^{\alpha+1}, \delta > 0.$$

D. *For* $f \in AC_1$,

$$\omega(f,\delta) \le \delta \int |f'|;$$

$$\omega^*(f,\delta) \le \delta\omega(f',\delta);$$

$$\omega^*(f,\delta) \le 2\delta \int |f'|, \delta > 0.$$

E. *For* $f \in AC_2$,

$$\omega^*(f,\delta) \le \delta^2 \int |f''|, \delta > 0.$$

We omit the proof of this simple Lemma. In Lemma 22, we will see how the second L_1 modulus of continuity can be used to obtain upper bounds

for $\int |S_m(f) - f|$. We will need the quantities

$$r_{mi} = 1 - \sqrt{\pi}\lambda_{mi}, \qquad i \geq 1.$$

LEMMA 22. *If K_m is an even kernel, then for all $f \in L_1[-\pi, \pi]$,*

$$\int |S_m(f) - f| \leq \int_0^\pi \omega^*(f, u)|K_m(u)|\, du.$$

If K_m is an even nonnegative kernel, and $f \in L_1[-\pi, \pi]$,

$$\int |S_m(f) - f| \leq A\omega^*\left(f, \sqrt{r_{m1}}\right),$$

where $A = \frac{1}{2}(1 + \pi/\sqrt{2})^2$ is a universal constant.

Proof. Note that

$$S_m(f) - f = \int_0^\pi (f(x+u) + f(x-u) - 2f(x))K_m(u)\, du,$$

so that

$$\int |S_m(f) - f| \leq \int_0^\pi \omega^*(f, u)|K_m(u)|\, du.$$

The second half of Lemma 22 requires some extra work. We have

$$\begin{aligned}
\int_{-\pi}^{\pi} u^2 K_m(u)\, du &\leq \int_{-\pi}^{\pi} \pi^2 \sin^2\left(\frac{u}{2}\right) K_m(u)\, du \\
&= \int_{-\pi}^{\pi} \frac{\pi^2}{2}(1 - \cos(u))K_m(u)\, du \\
&= \frac{\pi^2}{2}\left(1 - \lambda_{m1}\sqrt{\pi}\right) \\
&= \frac{\pi^2}{2} r_{m1},
\end{aligned}$$

and

$$\int_{-\pi}^{\pi} |u| K_m(u)\, du \le \left(\int_{-\pi}^{\pi} u^2 K_m(u)\, du \int_{-\pi}^{\pi} |K_m| \right)^{1/2} \le \frac{\pi}{\sqrt{2}} \sqrt{r_{m1}}\,.$$

If we use the fact that $\omega(f, t\delta) \le (1 + t)\omega(f, \delta)$, $\omega^*(f, t\delta) \le (1 + t)^2 \omega^*(f, \delta)$, $t, \delta > 0$, then we have

$$\int |S_m(f) - f| \le \omega^*\left(f, \frac{1}{t}\right) \int_0^{\pi} (1 + tu)^2 K_m(u)\, du$$

$$\le \omega^*\left(f, \frac{1}{t}\right)\left(\frac{1}{2} + 2t \int_0^{\pi} u K_m(u)\, du + t^2 \int_0^{\pi} u^2 K_m(u)\, du\right)$$

$$\le \omega^*\left(f, \frac{1}{t}\right)\left(\frac{1}{2} + 2t \frac{1}{2} \frac{\pi}{\sqrt{2}} \sqrt{r_{m1}} + t^2 \frac{\pi^2}{4} r_{m1}\right).$$

Substitute the value $1/t = \sqrt{r_{m1}}$.

For nonnegative kernels, r_{m1} gives us a clue about the goodness of the kernel: smaller values for r_{m1} mean a smaller bias under the same continuity conditions on f. Consider a few of the kernels we have described above:

(i) Fejér's kernel: $r_{m1} = 1/(m + 1)$.

(ii) Jackson's kernel: $r_{m1} = 3/(2m^2 + 1)$.

(iii) The Fejér–Korovkin kernel: $r_{m1} = 1 - \cos(\pi/(m + 2)) \le \pi^2/2m^2$ (and $r_{m1} \sim \pi^2/2m^2$).

(iv) de la Vallée Poussin's kernel: $r_{m1} = 1/(m + 1)$.

(v) Rogosinski's kernel: $r_{m1} \sim \pi^2/8m^2$.

(vi) de la Vallée Poussin's second kernel: $r_{m1} = 0$.

Kernels (v) and (vi) take negative values, and should thus not be compared on the basis of the second statement of Lemma 22. Based on Lemma 22 and these values of r_{m1}, the Jackson and Fejér–Korovkin kernels are more powerful than the Fejér and de la Vallée Poussin kernels. A combination of Lemmas 21 and 22 gives the following explicit bounds:

THEOREM 9 (Bounds for the Bias). *Let $K_m \ge 0$ be an even kernel, and let $r_{m1} \le C_1/m^p$ for some C_1, $p > 0$. If A is the constant of Lemma* 21, *then*

$\int |S_m(f) - f|$ *does not exceed the following quantities*:

$$A4\pi CC_1^{\alpha/2} m^{-p\alpha/2}, f \in W(0, \alpha, C);$$

$$A2\pi CC_1^{(\alpha+1)/2} m^{-p(\alpha+1)/2}, f \in W(1, \alpha, C);$$

$$A2\sqrt{C_1} \int |f'| m^{-p/2}, f \in AC_1;$$

$$AC_1 \int |f''| m^{-p}, f \in AC_2.$$

If we combine Theorem 9 with the bounds of either Theorem 8 or Lemma 20, we see that for some choice of m, the Jackson singular integral estimate and the Fejér–Korovkin singular integral estimate satisfy

$$E\left(\int |f_n - f|\right) = O(n^{-(s+\alpha)/(1+2(s+\alpha))}),$$

$\alpha \in (0, 1]$, $s = 0$ or $s = 1$. $E(\int |f_n - f|)$ is in fact uniformly bounded from above over these Lipschitz classes by the minimax lower bound for these classes times a constant not depending upon C or n. Thus, these estimates behave as the kernel estimate with nonnegative kernel. Also, for individual f, the asymptotic behavior of the bounds in Lemma 20 and Theorem 9 is similar to that of nonnegative kernel estimates, for example, the dependence upon $\int \sqrt{f}$, $\int |f'|$ ($f \in AC_1$) and $\int |f''|$ ($f \in AC_2$) is the same. Unfortunately, when $K_m \geq 0$, we encounter the same limitations as for the kernel estimate. This follows from the following Lemma taken from Butzer and Nessel (1971):

LEMMA 23 (Limitations of the Singular Integral Estimates). *Let K_m be a kernel satisfying the following properties*:

(i) $\lambda_{mi} = 0, i > m$ (*this can be replaced by* $i > cm$, *some* $c > 0$)

(ii) $\liminf_{m \to \infty} m^p r_{mk} > 0$, *some* $p > 0$, *all* $k \geq 1$.

Then, if f is a density on $[-\pi, \pi]$,

$$\liminf_{m \to \infty} m^p \int |S_m(f) - f| = 0$$

implies that $f(x) = 1/2\pi$, *almost everywhere*, $|x| \leq \pi$. *If K_m satisfies (i) and is nonnegative, then for all functions* $f \in L_1[-\pi, \pi]$ *not almost every-*

where equal to a constant function,

$$\liminf_{m\to\infty} m^2 \int |S_m(f) - f| > 0.$$

Proof. Note that $S_m(f)$ is a trigonometric polynomial of degree at most m (i.e., it is a linear combination of $\sin(kx), \cos(kx), k \le m$). Also,

$$\int_{-\pi}^{\pi} S_m(f,x)\frac{\cos(kx)}{\sqrt{\pi}}\,dx = \begin{cases} \sqrt{\pi}\lambda_{mk}a_{2k-1}, & 1 \le k \le m, \\ 0, & k > m; \end{cases}$$

$$\int_{-\pi}^{\pi} S_m(f,x)\frac{\sin(kx)}{\sqrt{\pi}}\,dx = \begin{cases} \sqrt{\pi}\lambda_{mk}a_{2k}, & 1 \le k \le m, \\ 0, & k > m. \end{cases}$$

Thus,

$$\int (S_m(f,x) - f(x))\left(\frac{\cos(kx) - i\sin(kx)}{\sqrt{\pi}}\right)dx$$

$$= (\sqrt{\pi}\lambda_{mk} - 1)(a_{2k-1} - ia_{2k}),$$

where i is the imaginary i. Therefore,

$$\frac{1}{\sqrt{\pi}}\int |S_m(f) - f| \ge |\sqrt{\pi}\lambda_{mk} - 1|\sqrt{a_{2k-1}^2 + a_{2k}^2}$$

$$= |r_{mk}|\sqrt{a_{2k-1}^2 + a_{2k}^2}.$$

Assume first that for all $k \ne 0$, $a_{2k-1}^2 + a_{2k}^2 = 0$. Then clearly, since the Fourier coefficients determine f uniquely, $S_m(f,x) = 1/2\pi$, all m, and $f(x) = 1/2\pi$, almost all x. If on the other hand $a_{2k-1}^2 + a_{2k}^2 > 0$ for a given $k \ne 0$, then for this k,

$$\liminf_{m\to\infty} \frac{\int |S_m(f) - f|}{|r_{mk}|} \ge \sqrt{\pi}\sqrt{a_{2k-1}^2 + a_{2k}^2} > 0,$$

which shows the first part of the lemma. For the second part, we use the Boas–Kac inequality (1945): if $\text{int}(\cdot)$ denotes the integer part of a real number, the inequality states that if K_m as in Lemma 23,

$$\lambda_{mk}\sqrt{\pi} \le \cos\left(\frac{\pi}{\text{int}(m/k) + 2}\right), \qquad 1 \le k \le m.$$

Thus,

$$\liminf_{m\to\infty} m^2|r_{mk}| \geq \liminf_{m\to\infty} m^2 \frac{\pi^2}{(\text{int}(m/k)+2)^2} \cdot \frac{1}{2} = \frac{\pi^2 k^2}{2} > 0.$$

We can now apply the first part of the lemma with $p = 2$.

Thus, for nonnegative K_m, no degree of smoothness imposed on f can help reduce $\int|S_m(f) - f|$ below $O(m^{-2})$. Thus, for these kernels, we cannot do better than the Jackson or Fejér–Korovkin kernels except perhaps by a constant factor. Our only hope for a reduced bias is a negative-valued kernel K_m. This will be further illustrated below.

Lemma 23 contains a lot of information about the best possible rates of convergence. In particular, for kernels $K_m \geq 0$ satisfying condition (i) of Lemma 23, we have

$$\liminf_{m\to\infty} m^2 \int|S_m(f) - f| \geq \sup_{k\geq 1} \sqrt{\pi}\, \frac{\pi^2 k^2}{2} \sqrt{a_{2k-1}^2 + a_{2k}^2},$$

and this is infinite whenever

$$\limsup_{k\to\infty} k^2|a_k| = \infty.$$

For the Fejér and de la Vallée Poussin kernels, and nonconstant f, the bias is bounded from below by a constant divided by m. Theorem 9 tells us this happens for $W(1,1,C)$ and AC_2, but for no other classes given there. In fact, the Fejér singular integral estimate attains an $O(1/m)$ bias for $f \in AC_2$ and for $f \in W(1,\alpha,C)$, all $\alpha > 0$. Thus, the argument given in Theorem 9 could give sub-optimal bounds in some cases. To obtain upper bounds of the right order, the first inequality of Lemma 22 can be used, as will be illustrated now on the Fejér singular integral.

LEMMA 24. *Let $S_m(f)$ be the Fejér singular integral for a density f on $[-\pi,\pi]$. Then,*

$$\int|S_m(f) - f| = \begin{cases} O(1/m), & f \in W(1,\alpha,C),\ \textit{all}\ \alpha \in (0,1],\ \textit{or}\ f \in AC_2; \\ O(\log m/m), & f \in W(0,1,C)\ \textit{or}\ f \in AC_1; \\ O(m^{-\alpha}), & f \in W(0,\alpha,C), \alpha \in (0,1). \end{cases}$$

Proof. In our argument, we will use a bounding argument of Butzer and Nessel (1971, p. 81): by Lemma 4, for $u > 0$,

$$2\pi u^{\alpha} F_m(u) \le \begin{cases} 2^{\alpha-1}\pi(m+1)^{1-\alpha}, & 0 < u \le 1/m; 0 < \alpha \le 1; \\ \pi^2 u^{\alpha-2}(m+1)^{-1}, & 0 < u \le \pi; 0 < \alpha \le 2. \end{cases}$$

Thus, $\int_0^{1/m} u^{\alpha} F_m(u)\,du \le m^{-\alpha}, 0 < \alpha \le 1$, and

$$\int_{1/m}^{\pi} u^{\alpha} F_m(u)\,du \le \begin{cases} \frac{\pi}{2}(1-\alpha)^{-1} m^{-\alpha}, & 0 < \alpha \le 1; \\ \frac{\pi}{2}\frac{\log(\pi m)}{m+1}, & \alpha = 1. \end{cases}$$

This gives the bounds

$$\int_0^{\pi} u^{\alpha} F_m(u)\,du \le \begin{cases} m^{-\alpha}\left(1 + \frac{\pi}{2(1-\alpha)}\right), & \alpha \in (0,1); \\ \frac{1}{m} + \frac{\pi}{2}\frac{\log(\pi m)}{m+1}, & \alpha = 1; \\ \frac{\pi^{\alpha}}{2(\alpha-1)}\frac{1}{m}, & \alpha \in (1,2]. \end{cases}$$

Lemma 24 follows if we combine these bounds with Lemma 21 and the first inequality of Lemma 22. (For all the classes of densities of interest to us, $\int |S_m(f) - f| \le C\int_0^{\pi} u^{\alpha} F_m(u)\,du$ for some constants C and α.)

Essentially, with the Fejér singular integral estimate, there is no hope of obtaining $E(\int |f_n - f|) = o(n^{-1/3})$ except for the constant density on $[-\pi, \pi]$. In this sense, the Fejér and de la Vallée Poussin singular integral estimates behave as the histogram estimate. At the other end of the scale are the singular integral estimates with unlimited power, that is, estimates with

$$\int |S_m(f) - f| = O(m^{-p})$$

for any power $p \ge \frac{1}{2}$ provided f is "smooth enough." From Lemma 23 we can see that necessarily K_m is negative-valued and for all $i \ne 0$, all $p > 0$,

$$\liminf_{m \to \infty} m^p r_{mi} = 0.$$

The latter condition is satisfied by de la Vallée Poussin's second kernel (because $r_{mi} = 0, 1 \le i \le m$) and by the Dirichlet kernel. Basically, the λ_{mi}'s must be flat near the origin ($\sqrt{\pi}\lambda_{mi} = 1$, all i smaller than a number diverging to infinity as $m \to \infty$). For such kernels, there is hope to obtain any rate of convergence of the bias without changing kernels along the way. For similar behavior, we refer to the trapezoidal kernel estimate.

It is worth pointing out that Rogosinski's negative-valued kernel satisfies (i) and (ii) of Lemma 23 with power $p = 2$. Thus, it has the same limitations as Jackson's kernel and positive-valued kernels. This is disappointing, for if we are sacrificing positivity in our density estimate f_n, we might as well choose a kernel with unlimited power as described above. Hall (1983) has compared Rogosinski's singular integral estimate with other estimates based on positive kernels, and found the same rate of convergence (in L_2) but a smaller constant. In a sense, that is "cheating". In fact, for unlimited power kernels, we will see that a better rate of convergence is obtainable under the *same* smoothness conditions on f.

We conclude this section with the description of the properties of one kernel with unlimited power. To obtain refined rates of convergence, we will use Jackson's first and second theorems. This technique provides the reader with another set of tools (recall that in Section 5, we used Lorentz's inequality to handle the trigonometric series estimate).

LEMMA 25 (Jackson's Theorems). *Let T_m be the class of all trigonometric polynomials of degree at most m, and let $f \in L_1[-\pi, \pi]$.*
Jackson's First Theorem

$$\inf_{t_m \in T_m} \int |t_m - f| \le 2A^2 \omega^*(f, 1/m),$$

where A is the constant of Lemma 22.
Jackson's Second Theorem

$$\inf_{t_m \in T_m} \int |t_m - f| \le \begin{cases} \left(\frac{36}{m}\right)^s \int |f^{(s)}|; \\ \frac{36^s(36^2+1)}{m^s} \omega^*\left(f^{(s)}, \frac{1}{m}\right), \end{cases}$$

for all $f \in AC_s$, all $s > 0$.

Proof. No attempt will be made to obtain the best possible constants (see, e.g., Butzer and Nessel (1971) for references dealing with the best possible constants, and for a complete proof of Jackson's second theorem). We will merely prove Jackson's first theorem.

We observe that for the Fejér–Korovkin kernel, $S_m(f) \in T_m$, and thus

$$\inf_{t_m \in T_m} \int |t_m - f| \leq \int |S_m(f) - f|.$$

Also,

$$\int |S_m(f) - f| \leq A\omega^*\left(f, \left(1 - \cos\left(\frac{\pi}{m+2}\right)\right)^{1/2}\right) \quad \text{(Lemma 22)}$$

$$\leq A\omega^*\left(f, \frac{\pi}{(m+2)\sqrt{2}}\right)$$

$$\leq A\left(1 + \frac{\pi}{\sqrt{2}}\right)^2 \omega^*\left(f, \frac{1}{m}\right)$$

$$= \frac{1}{2}\left(1 + \frac{\pi}{\sqrt{2}}\right)^4 \omega^*\left(f, \frac{1}{m}\right).$$

This concludes the proof of Jackson's first theorem.

THEOREM 10 (The Singular Integral Estimate with de la Vallée Poussin's Second Kernel.) *Let f_n be the singular integral estimate with de la Vallée Poussin's second kernel (denoted here by K_m), and let $S_m(f)$ be the corresponding singular integral for a density f on $[-\pi, \pi]$.*

A. *K_m is an approximate identity, and thus*

$$\int |S_m(f) - f| = o(1), \text{ all } f.$$

B. $\int |S_m(g)| \leq 3 \int |g|$, *all* $g \in L_1[-\pi, \pi]$.

C. $\int |S_m(f) - f| \leq 4 \inf_{t_m \in T_m} \int |t_m - f|$.

D. $\int K_m^2/n \leq 9m/4n$.

E. *$K_m^2/\int K_m^2$ is an approximate identity.*

F. *If* $\lim_{n\to\infty} m = \infty$, $\lim_{n\to\infty}(m/n) = 0$, *then*

$$E\left(\int |f_n - f|\right) = o(1), \text{ all } f.$$

G. $$E\left(\int|f_n - f|\right) \le 4 \inf_{t_m \in T_m} \int|t_m - f| + \begin{cases} \left(\int\sqrt{f} + o(1)\right)\sqrt{9m/4n} \\ \sqrt{9\pi m/2n}\,. \end{cases}$$

H. *If M is a constant integer, $f \in T_M$, and m is constant, $m \ge M$, then* $\int|S_m(f) - f| = 0$, *and*

$$E(|f_n - f|) = O(n^{-1/2}).$$

I. *If $f \in AC_s$, $s > 0$ integer, then*

$$\inf_m E\left(\int|f_n - f|\right) = o(n^{-s/(2s+1)}).$$

J. *If $f \in AC_s$, $s > 0$ integer, then*

$$E\left(\int|f_n - f|\right) \le 4\left(\frac{36}{m}\right)^s \int|f^{(s)}| + \begin{cases} \left(\int\sqrt{f} + o(1)\right)\sqrt{9m/4n} \\ \sqrt{9\pi m/2n}\,. \end{cases}$$

K. *If $f \in W(s, \alpha, C)$ for some integer $s \ge 0, \alpha \in (0, 1]$, $C > 0$, then*

$$E\left(\int|f_n - f|\right) \le 4(36)^s(36^2 + 1)\frac{4\pi C}{m^{s+\alpha}} + \sqrt{\frac{9\pi m}{2n}}$$

and

$$\inf_m E\left(\int|f_n - f|\right) = O(n^{-(s+\alpha)/(2(s+\alpha)+1)}).$$

Proof. A follows from Lemma 19. For B, we note that

$$\int|K_m| \le \int|(1 + 2\cos(mu))|F_{m-1}(u)\,du \le 3\int F_{m-1}(u)\,du = 3,$$

and thus that

$$\int|S_m(g)| \le \int|g|\int K_m| \le 3\int|g|.$$

For $t_m \in T_m$, we have $S_m(t_m) = t_m$. Thus,

$$\int |S_m(f) - f| \le \inf_{t_m \in T_m} \left(\int |S_m(f) - S_m(t_m)| + \int |t_m - f| \right)$$

$$\le 4 \inf_{t_m \in T_m} \int |t_m - f|,$$

which proves C. D follows from the crude bounds of Lemma 4:

$$\int K_m^2 \le \int 9F_{m-1}^2 \le 9 \sup F_{m-1} \le \frac{9m}{4}.$$

Property E is trivial. The consistency (property F) follows from properties A and D, and Theorem 8. Properties C and D and the inequalities of Theorem 8 and Lemma 20 give us property G. Property H follows directly from this. If we use the fact that $\omega^*(f^{(s)}, 1/m) = o(1)$ for $f \in AC_s$, then property I follows from property G and the last inequality in Jackson's second theorem. Using the first inequality in Jackson's second theorem gives us property J. Finally, property K can be deduced from property G, Jackson's second theorem and part B of Lemma 21:

The singular integral estimate of Theorem 10 is only a slight modification of the trigonometric series estimate; yet it is safer to use because it is universally consistent. Also, the estimate on the bias (property C) improves over the corresponding estimate for the trigonometric series estimate (Theorem 4) by a factor of $\log m$. Just as the trigonometric series estimate or the trapezoidal kernel estimate, this estimate has an expected L_1 error that comes to within a constant of the minimax lower bound for $W(s, \alpha, C)$, *all* s (see property K). This is why we could call this an estimate with unlimited power. Even the rate $O(n^{-1/2})$ is attainable, coupled with unbiasedness for all n (property H). Finally, property I shows how the estimate improves over all singular integral estimates with $K_m \ge 0$ and Rogosinski's estimate, even for AC_2: for the latter estimates, all we could hope for is an error of size $O(n^{-2/5})$ since the bias must at least be of the order of m^{-2} (Lemma 23).

REFERENCES

G. L. Anderson and R. J. P. de Figueiredo (1980). An adaptive orthogonal-series estimator for probability density functions, *Annals of Statistics* **8**, pp. 347—376.

R. Askey and S. Wainger (1965). Mean convergence of expansions in Laguerre and Hermite series, *American Journal of Mathematics* **87**, pp. 695–708.

J. P. Asselin de Beauville (1978). Estimation de la densité de probabilité par une série de polynômes d'Hermite. Détermination du nombre optimal de termes de la série, *Comptes Rendus de l'Académie des Sciences de Paris* **286**, pp. 309–311.

N. K. Bary (1964a). *A Treatise on Trigonometric Series*, Vol. 1, Pergamon Press, Oxford.

N. K. Bary (1964b). *A Treatise on Trigonometric Series*, Vol. 2, Pergamon Press, Oxford.

J. Bleuez and D. Bosq (1976). Conditions nécessaires et suffisantes de convergence pour une classe d'estimateurs de la densité, *Comptes Rendus de l'Académie des Sciences de Paris* **282**, pp. 63–66.

J. Bleuez and D. Bosq (1979). Conditions nécessaires et suffisantes de convergence de l'estimateur de la densité par la méthode des fonctions orthogonales, *Revue Roumaine de Mathématiques Pûres et Appliquées* **24**, pp. 869–886.

R. P. Boas and M. Kac (1945). Inequalities for Fourier transforms of positive functions, *Duke Mathematical Journal* **12**, pp. 189–206.

D. Bosq (1969). Sur l'estimation d'une densité multivariée par une série de fonctions orthogonales, *Comptes Rendus de l'Académie des Sciences de Paris* **268**, pp. 555–557.

D. Bosq and J. Bleuez (1978). Etude d'une classe d'estimateurs non-paramétriques de la densité, *Annales de l'Institut Henri Poincaré* **14**, pp. 479–498.

H. D. Brunk (1977). Univariate density estimation by orthogonal series, with application to estimation of wildlife populations by line transect surveys, unpublished manuscript.

H. B. Brunk (1978). Univariate density estimation by orthogonal series, *Biometrika* **65**, pp. 521–528.

P. L. Butzer and R. J. Nessel (1971). *Fourier Analysis and Approximation*, Vol. 1, Birkhäuser Verlag, Basel and Stuttgart.

L. Carleson (1966). On convergence and growth of partial sums of Fourier series, *Acta Mathematica* **116**, pp. 135–157.

N. N. Cenov (1962). Evaluation of an unknown distribution density from observations, *Soviet Mathematics* **3**, pp. 1559–1562.

B. R. Crain (1973). A note on density estimation using orthogonal expansions, *Journal of the American Statistical Association* **68**, pp. 964–965.

B. R. Crain (1974). Estimation of distributions using orthogonal expansions, *Annals of Statistics* **2**, pp. 454–463.

R. E. Edwards (1979). *Fourier Series. A Modern Introduction*, Vol. 1, Springer-Verlag, Berlin.

W. H. Fellner (1974). Heuristic estimation of probability densities, *Biometrika* **61**, pp. 485–492.

C. Fefferman (1971). On the convergence of multiple Fourier series, *Bulletin of the American Mathematical Society* **77**, pp. 744–755.

A. Földes and P. Révész (1974). A general method for density estimation, *Studia Scientiarium Mathematicarum Hungarica* **9**, pp. 81–92.

W. Greblicki (1981). Asymptotical efficiency of classifying procedures using the Hermite series estimate of multivariate probability densities, *IEEE Transactions on Information Theory* **IT-27**, pp. 364–366.

W. Greblicki and M. Pawlak (1981). Classification using the Fourier series estimate of multivariate density functions, *IEEE Transactions on Systems, Man and Cybernetics* **SMC-11**, pp. 726–730.

P. Hall (1981). On trigonometric series estimates of densities, *Annals of Statistics* **9**, pp. 683–685.

P. Hall (1982). Comparison of two orthogonal series methods of estimating a density and its derivatives on an interval, *Journal of Multivariate Analysis* **12**, pp. 432–449.

P. Hall (1983). Measuring the efficiency of trigonometric series estimates of a density, *Journal of Multivariate Analysis* **13**, pp. 234–256.

R. A. Hunt (1968). On the convergence of Fourier series. Orthogonal expansions and their continuous analogues, Proceedings of a Conference held at Edwardsville, Illinois, 1967. Southern Illinois University Press, Carbondale, pp. 235–255.

O. G. Jorsboe and L. Mejlbro (1982). *The Carleson–Hunt theorem of Fourier Series*, Springer-Verlag, Berlin.

A. N. Kolmogorov (1926). Une série de Fourier–Lebesgue divergente partout, *Comptes Rendus de l'Académie des Sciences de Paris* **183**, pp. 1327–1328.

T. W. Körner (1981). Everywhere divergent Fourier series, *Colloquium in Mathematics* **45**, pp. 103–118.

R. Kronmal and M. Tarter (1968). The estimation of probability densities and cumulatives by Fourier series methods, *Journal of the American Statistical Association* **63**, pp. 925–952.

A. Krzyzak and M. Pawlak (1982). Estimation of a multivariate density by orthogonal series, in *Probability and Statistical Inference*, W. Grossmann et al. (Eds.), Reidel, Hingham, MA, pp. 211–221.

G. G. Lorentz (1948). Fourier–Koeffizienten und Funktionenklassen, *Mathematische Zeitschrift* **51**, pp. 135–149.

C. J. Mozzochi (1971). *On the Pointwise Convergence of Fourier Series*, Springer-Verlag, Berlin.

B. Muckenhoupt (1969). Mean convergence of Jacobi series, *Proceedings of the American Mathematical Society* **23**, pp. 306–310.

B. Muckenhoupt (1970a). Equiconvergence and almost everywhere convergence of Hermite and Laguerre series, *SIAM Journal of Mathematical Analysis* **1**, pp. 295–321.

B. Muckenhoupt (1970b). Mean convergence of Hermite and Laguerre series. I, *Transactions of the American Mathematical Society* **147**, pp. 419–431.

B. Muckenhoupt (1970c). Mean convergence of Hermite and Laguerre series. II, *Transactions of the American Mathematical Society* **147**, pp. 433–460.

J. Newman and W. Rudin, (1952). Mean convergence of orthogonal series, *Proceedings of the American Mathematical Society* **3**, pp. 219–222.

A. M. Olevskii (1975). *Fourier Series with Respect to General Orthogonal Systems*, Springer-Verlag, Berlin.

H. Pollard (1947). The mean convergence of orthogonal series. I, *Transactions of the American Mathematical Society* **62**, pp. 387–403.

H. Pollard (1948). The mean convergence of orthogonal series. II, *Transactions of the American Mathematical Society* **63**, pp. 355–367.

H. Pollard (1949). The mean convergence of orthogonal series. III, *Duke Mathematical Journal* **16**, pp. 189–191.

E. S. Quade (1937). Trigonometric approximation in the mean, *Duke Mathematical Journal* **3**, pp. 529–543.

M. Rosenblatt (1971). Curve estimates, *Annals of Mathematical Statistics* **42**, pp. 1815–1841.

G. Sansone (1977). *Orthogonal Functions*, Krieger, Huntington, NY.

L. Schüler (1976). Über die Konsistenz einer Schätzung mehrdimensionaler Dichten auf der Basis trigonometrischer Reihen, *Metrika* **23**, pp. 77–82.

S. C. Schwartz (1967). Estimation of a probability density by an orthogonal series, *Annals of Mathematical Statistics* **38**, pp. 1262–1265.

P. Sjölin (1971). Convergence almost everywhere of certain singular integrals and multiple Fourier series, *Arkiv für Mathematik* **9**, pp. 65–90.

H. Stegbuchner (1980). Dichteschätzungen mit Gleichverteilungsmethoden, *Periodica Mathematica Hungarica* **11**, pp. 161–175.

E. M. Stein (1961). On limits of sequences of operators, *Annals of Mathematics* **74**, pp. 140–170.

H. Sterbuchner (1980). On nonparametric multivariate density estimation, *Revue Roumaine de Mathématiques Pûres et Appliquées* **25**, pp. 111–118.

G. Szegö (1975). *Orthogonal Polynomials*, Vol. 23, 4th Ed., American Mathematical Society Colloquia Publications, Providence, RI.

M. E. Tarter and R. A. Kronmal (1976). An introduction to the implementation and theory of nonparametric density estimation, *The American Statistician* **30**, pp. 105–112.

J. Van Ryzin (1966). Bayes risk consistency of classification procedures using density estimation, *Sankhya Series A* **28**, pp. 261–270.

A. J. Viollaz (1980). Asymptotic distribution of L_2 norms of the deviations of density function estimates, *Annals of Statistics* **8**, pp. 322–346.

G. Wahba (1975). Optimal convergence properties of variable knot, kernel, and orthogonal series methods for density estimation, *Annals of Statistics* **3**, pp. 15–29.

G. Wahba (1977). Optimal smoothing of density estimates, in *Classification and Clustering*, J. Van Ryzin (Ed.) Academic Press, New York, pp. 423–458.

G. Wahba (1978). Data-based optimal smoothing of orthogonal series density estimates, Department of Statistics, University of Wisconsin, Technical Report 509.

G. G. Walter (1977). Properties of Hermite series estimation of probability density, *Annals of Statistics* **5**, pp. 1258–1264.

G. G. Walter and J. R. Blum (1979). Probability density estimation using delta sequences, *Annals of Statistics* **7**, pp. 328–340.

G. S. Watson (1969). Density estimation by orthogonal series, *Annals of Mathematical Statistics* **40**, pp. 1496–1498.

P. Whittle (1958). On the smoothing of probability density functions, *Journal of the Royal Statistical Society B* **20**, pp. 334–343.

A. Zygmund (1959). *Trigonometric Series*, **Vols. 1, 2**, Cambridge University Press, Cambridge, U.K.

Author Index

Numbers in *italics* refer to the pages on which the complete references are listed.

Subject Index

Numbers in *italics* refer to implicit or explicit definitions, or to key results.

Applied Probability and Statistics (Continued)

DODGE and ROMIG • Sampling Inspection Tables, *Second Edition*
DOWDY and WEARDEN • Statistics for Research
DRAPER and SMITH • Applied Regression Analysis, *Second Edition*
DUNN • Basic Statistics: A Primer for the Biomedical Sciences, *Second Edition*
DUNN and CLARK • Applied Statistics: Analysis of Variance and Regression
ELANDT-JOHNSON and JOHNSON • Survival Models and Data Analysis
FLEISS • Statistical Methods for Rates and Proportions, *Second Edition*
FOX • Linear Statistical Models and Related Methods
FRANKEN, KÖNIG, ARNDT, and SCHMIDT • Queues and Point Processes
GALAMBOS • The Asymptotic Theory of Extreme Order Statistics
GIBBONS, OLKIN, and SOBEL • Selecting and Ordering Populations: A New Statistical Methodology
GNANADESIKAN • Methods for Statistical Data Analysis of Multivariate Observations
GOLDBERGER • Econometric Theory
GOLDSTEIN and DILLON • Discrete Discriminant Analysis
GREENBERG and WEBSTER • Advanced Econometrics: A Bridge to the Literature
GROSS and CLARK • Survival Distributions: Reliability Applications in the Biomedical Sciences
GROSS and HARRIS • Fundamentals of Queueing Theory
GUPTA and PANCHAPAKESAN • Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations
GUTTMAN, WILKS, and HUNTER • Introductory Engineering Statistics, *Third Edition*
HAHN and SHAPIRO • Statistical Models in Engineering
HALD • Statistical Tables and Formulas
HALD • Statistical Theory with Engineering Applications
HAND • Discrimination and Classification
HILDEBRAND, LAING, and ROSENTHAL • Prediction Analysis of Cross Classifications
HOAGLIN, MOSTELLER, and TUKEY • Understanding Robust and Exploratory Data Analysis
HOEL • Elementary Statistics, *Fourth Edition*
HOEL and JESSEN • Basic Statistics for Business and Economics, *Third Edition*
HOGG and KLUGMAN • Loss Distributions
HOLLANDER and WOLFE • Nonparametric Statistical Methods
IMAN and CONOVER • Modern Business Statistics
JAGERS • Branching Processes with Biological Applications
JESSEN • Statistical Survey Techniques
JOHNSON and KOTZ • Distributions in Statistics
Discrete Distributions
Continuous Univariate Distributions—1
Continuous Univariate Distributions 2
Continuous Multivariate Distributions
JOHNSON and KOTZ • Urn Models and Their Application: An Approach to Modern Discrete Probability Theory
JOHNSON and LEONE • Statistics and Experimental Design in Engineering and the Physical Sciences, Volumes I and II, *Second Edition*
JUDGE, HILL, GRIFFITHS, LÜTKEPOHL and LEE • Introduction to the Theory and Practice of Econometrics
JUDGE, GRIFFITHS, HILL, LÜTKEPOHL and LEE • The Theory and Practice of Econometrics, *Second Edition*
KALBFLEISCH and PRENTICE • The Statistical Analysis of Failure Time Data
KEENEY and RAIFFA • Decisions with Multiple Objectives

Applied Probability and Statistics (Continued)

LAWLESS • Statistical Models and Methods for Lifetime Data
LEAMER • Specification Searches: Ad Hoc Inference with Nonexperimental Data
LEBART, MORINEAU, and WARWICK • Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices
McNEIL • Interactive Data Analysis
MAINDONALD • Statistical Computation
MANN, SCHAFER and SINGPURWALLA • Methods for Statistical Analysis of Reliability and Life Data
MARTZ and WALLER • Bayesian Reliability Analysis
MIKÉ and STANLEY • Statistics in Medical Research: Methods and Issues with Applications in Cancer Research
MILLER Beyond ANOVA: • Basics of Applied Statistics
MILLER • Survival Analysis
MILLER, EFRON, BROWN, and MOSES • Biostatistics Casebook
MONTGOMERY and PECK • Introduction to Linear Regression Analysis
NELSON • Applied Life Data Analysis
OTNES and ENOCHSON • Applied Time Series Analysis: Volume I, Basic Techniques
OTNES and ENOCHSON • Digital Time Series Analysis
PANKRATZ • Forecasting with Univariate Box-Jenkins Models: Concepts and Cases
PIELOU • Interpretation of Ecological Data: A Primer on Classification and Ordination
POLLOCK • The Algebra of Econometrics
PRENTER • Splines and Variational Methods
RAO and MITRA • Generalized Inverse of Matrices and Its Applications
RIPLEY • Spatial Statistics
SCHUSS • Theory and Applications of Stochastic Differential Equations
SEAL • Survival Probabilities: The Goal of Risk Theory
SEARLE • Linear Models
SEARLE • Matrix Algebra Useful for Statistics
SPRINGER • The Algebra of Random Variables
STOYAN • Comparison Methods for Queues and Other Stochastic Models
UPTON • The Analysis of Cross-Tabulated Data
WEISBERG • Applied Linear Regression
WHITTLE • Optimization Over Time: Dynamic Programming and Stochastic Control, Volume I and Volume II
WILLIAMS • A Sampler on Sampling
WONNACOTT and WONNACOTT • Econometrics, *Second Edition*
WONNACOTT and WONNACOTT • Introductory Statistics, *Fourth Edition*
WONNACOTT and WONNACOTT • Introductory Statistics for Business and Economics, *Third Edition*
WONNACOTT and WONNACOTT • Regression: A Second Course in Statistics
WONNACOTT and WONNACOTT • Statistics: Discovering Its Power
ZELLNER • An Introduction to Bayesian Inference in Econometrics

Tracts on Probability and Statistics

AMBARTZUMIAN • Combinatorial Integral Geometry
BIBBY and TOUTENBURG • Prediction and Improved Estimation in Linear Models
BILLINGSLEY • Convergence of Probability Measures
DEVROYE and GYORFI • Nonparametric Density Estimation: The L_1 View
KELLY • Reversibility and Stochastic Networks
RAKTOE, HEDAYAT, and FEDERER • Factorial Designs
TOUTENBURG • Prior Information in Linear Models