Progress in Probability
and Statistics

Series Editor
Murray Rosenblatt

Luc Devroye

# A Course in Density Estimation

**Progress in Probability and Statistics**
Volume 14

Series Editor
Murray Rosenblatt

Luc Devroye

# A Course in Density Estimation

1987

Birkhäuser
Boston · Basel · Stuttgart

Luc Devroye
School of Computer Science
McGill University
Montreal H3A 2K6
Canada

Library of Congress Cataloging in Publication Data
Devroye, Luc.
A course in density estimation.
(Progress in probability and statistics ; v. 14)
Bibliography: p.
Includes index.
1. Estimation theory. I. Title. II. Title: Density estimation. III. Series.
QA276.8.D48 1987 519.5'44 87–829

CIP-Kurztitelaufnahme der Deutschen Bibliothek
Devroye, Luc:
A course in density estimation / Luc Devroye.—
Boston ; Basel ; Stuttgart : Birkhäuser, 1987.
—195 S.
(Progress in probability and statistics ; Vol. 14)
ISBN 3-7643-3365-0 (Basel . . .)
ISBN 0-8176-3365-0 (Boston)
NE: GT



ISBN 0-8176-3365-0
ISBN 3-7643-3365-0

Printed and bound by R.R. Donnelley & Sons, Harrisonburg, Virginia.
Printed in the U.S.A.

9 8 7 6 5 4 3 2 1

# TABLE OF CONTENTS

# PREFACE

These notes were written during the summer quarter of 1986, when I taught a course on density estimation in the Department of Statistics at Stanford University. I am grateful to the students (Naomi Altman, Tony Cooper, Joseph Gagnon, Eric Holmgren, Yoichi Ii, Stephen Langlois and Jia Yang Sun) and to the local and visiting faculty (in particular Will Gersch, Petter Laake, Art Owen, David Scott and Hermann Thorisson) for invaluable feedback.

The original manuscript was entitled **A Summer Course On Density Estimation**, to reflect the warm relaxed atmosphere in which the book was written, and to indicate the potential market as a textbook for a summer quarter. In this preface, we will explain how the material of this book hangs together, how the text is related to Devroye and Gyorfi (1985), where and how density estimates are applied, and where more theoretical research is needed in the area.

**The contents.**

An iid sequence of $n$ random variables with a uniform distribution on the surface of the unit sphere of $R^3$ has the interesting property that the $n$ $x$-coordinates (projections) form an iid sequence with a uniform distribution on $[-1,1]$. In general, it is possible to define random vectors with a uniform distribution such that their projections have a given density $f$ : just consider $n$ iid random variables uniformly distributed under the curve of $f$ (i.e., on $\{(x,y) : 0 \leq y \leq f(x), x \in R\}$); their projections form an iid sequence drawn from $f$.

In density estimation, we are only presented with the projections, and are asked to reconstruct, or estimate, $f$. Usually, the density estimate itself is a density too, i.e. it is nonnegative and integrates to one.

First of all, it is necessary to pick a criterion for judging the goodness of an estimate. This is perhaps the most critical stage in the entire undertaking. In the context of the projections discussed above, we can easily construct such a criterion: consider for example the minimal area we need to take away from the estimate $f_n$ and give (paste) to the estimated density $f$. The relocated area is

$\int |f_n - f|/2$, which is also known as the total variation. The $L_1$ error, $\int |f_n - f|$, is a number between 0 and 2. Interestingly, if we transform the space by any one-to-one onto transformation, the $L_1$ error is unaffected. In other words, it is a universal measure of the closeness of $f_n$ to $f$. Some other distances, such as the $L_p$ distance for $p \neq 1$, are not even invariant under a simple rescaling of the axes. If $f_n$ induces probability measure $\mu_n$, and $f$ induces $\mu$, then it is not difficult to see that the $L_1$ error is $2 \sup_A |\mu_n(A) - \mu(A)|$ where the supremum is over all (Borel) sets $A$. This interpretation in terms of differences between probabilities makes the $L_1$ criterion unique. When someone in the field reports an error of 0.012, then we know that all probabilities of all sets are off by at most 0.006. $L_1$ errors are not only easily interpreted, they are also easily visualized: the visual impression of the distance between the plots of $f_n$ and $f$ is precisely the area between the curves, $\int |f_n - f|$. The visual aspect is of course increasingly important in view of the invasion of graphical tools for presenting one's results on workstations, terminal screens and laser printers. One often finds technical reports and papers in which plots of densities are shown besides tables of distances between those densities, where the distance is not the $L_1$ distance. In many cases, there is no continuous relationship between spaces of densities that are endowed with different distances, so that the plots do not necessarily show close densities when the numbers in the tables are small, and vice versa. In **chapter I**, we will relate the $L_1$ distance to other distances between densities, such as $L_p$ distances, Hellinger distances and distances based upon Kullback-Leibler numbers. Often one can deduce results indirectly by using an appropriate inequality between distance measures. The existence of such inequalities is also discussed in chapter I. One important point to note here is that the $L_1$ distance is just one of many distances that one can define on the space of **all** densities. This, of course, will lead to a theory that is uncluttered by hard-to-verify conditions.

Our criterion is global since most of the interesting applications demand global goodness of one or more density estimates. Local criteria, such as pointwise convergence, are much less important, since they are oblivious to the role of a density as a probability measure: it is possible to construct estimates that are very good at one or more points, but integrate to infinity.

In **chapter II**, we introduce the density estimation problem, and derive several density estimates, based upon the fact that a density $f$, the Radon-Nikodym derivative of a probability measure $\mu$ with respect to Lebesgue measure, can be approximated by a ratio of an empirical measure of a small ball to Lebesgue measure of the same ball. Estimates constructed in this fashion are designed to work for all $f$, and will be called nonparametric. "Universal" estimates would have been a better term, but it would be silly to replace a term that has been in use for over 20 years. It is possible to tailor-design estimates for particular classes of densities; for example, a beta density can be estimated by another beta density in which the two beta parameters are data-based estimates of the unknown beta parameters. The risk of this approach is that if the assumption that the data are beta distributed is false, then all is lost, for there is no hope of

approaching the underlying density. But the benefits can be sweet, for if the assumption is correct, we will latch on to the beta density, and obtain very good estimates. Estimates in which a finite number of parameters are estimated in an otherwise inflexible model are called parametric. Tailor-designed estimates on the other hand are estimates that are designed to perform well for a proper subclass ("target class") of all densities; they can but don't have to be parametric.

In the notes, we will concentrate on nonparametric and tailor-designed estimates. We will deal with some parametric estimates in the exercises. We mainly study the kernel estimate, first introduced by Parzen and Rosenblatt. This estimate is nothing but an equiprobable mixture of $n$ similar-shaped densities (kernels) centered at the data points. It easy to understand, analyze, and modify. The flexibility comes from our ability to pick the shape of the kernel, possibly as a function of the data. Unfortunately, we can't show here that the kernel estimate's performance matches that of most other estimates for most densities, since that would force us to introduce other estimates to find out later that we didn't need them in the first place. The reader will just have to trust our judgment. Note however that the kernel estimate is not a cure-all: we will find out that it does a fairly good job for most densities, but that it can't compete with tailor-designed estimates for specific small target classes of densities, such as the class of all normal densities, or the class of all mixtures of two beta densities.

In **chapter III**, we show that the kernel estimate is guaranteed to converge to $f$ if we choose its parameters in a certain way. Since $\int |f_n - f|$ is a random variable, we can study many modes of convergence. However, the story is quite simple, since all modes of convergence are equivalent to each other. Furthermore, there is no density for which the kernel estimate is not consistent. This is due to the fact that the collection of nice densities (where "nice" could mean bounded, of compact support, and infinitely many times continuously differentiable) is dense in the $L_1$ space of all densities, so that each pathological density is surrounded by infinitely many nice densities. In fact, a sample of size $n$ drawn from $f$ can be thought of as a sample drawn from a nice neighboring density $g$ if we are allowed to replace a few (about $\frac{n}{2}\int |f-g|$) data points by other ones. Close densities require very little sample surgery.

There are many other issues that matter when one estimates a density. Some of these are sprinkled throughout the text, without any strong sense of order. Chapters IV, V and VI deal with universal properties of all density estimates, and introduce notions that can aid users in deciding which density estimate they should choose for the problem at hand.

For example, in **chapter IV**, we define robustness for density estimates in terms of $L_1$ distances, and verify that many estimates we think of intuitively as non-robust (or hyper-sensitive) are indeed declared non-robust by our robustness criterion. The kernel and histogram estimates are very robust, since the removal or replacement of one data point hardly affects the overall estimate, its weight in the kernel mixture being only $1/n$. Robustness is not a selection criterion per se. The main point of chapter IV is to show that robustness is equivalent to insensitivity to small surgery on the sample.

The main selection criterion is of course $\int |f_n - f|$ itself. Unfortunately, this integral is a random variable, as it is a function of the data. Should we compare means, medians, 90 percentiles, or mean-plus-standard-deviation? Pessimists, always assuming the worst possible scenario, would consider ess sup$\int |f_n - f|$ (where ess sup is the essential supremum). This is nearly always 2, since we take in fact the essential supremum with respect to all possible samples of size $n$. Gamblers may want to argue on the basis of the mean, taking the risk that when the standard deviation is much larger than the mean, $f_n$ behaves erratically. If $\int |f_n - f|$ has mean $J_n$ and variance $\sigma_n^2$, it is not difficult to show that $\sigma_n \leq \sqrt{2J_n - J_n^2}$, where equality can be attained. Since $J_n \to 0$ in the cases of interest to us, it is possible to have $(J_n + \sigma_n)/J_n \to \infty$. Thus, $J_n + \sigma_n$ and $J_n$ can tend to zero at different rates. The same is true for $J_n$ and $m_n$, the median of $\int |f_n - f|$, since we can have $m_n = 0$ for any value $J_n \leq 1$. In other words, we are faced with a crucial decision about the choice of a deterministic number that is representative of the "usual values" of $\int |f_n - f|$.

Luckily, most nonparametric estimates such as the kernel estimate and histogram estimate are relatively stable, i.e. $\int |f_n - f| / E(\int |f_n - f|) \to 1$ in probability. For example, they satisfy the property that $\sigma_n / J_n \to 0$ as $n \to \infty$. This law of large numbers is proved in **chapter IX**, where more information is given about the closeness to the limit in terms of exponential inequalities. It can be explained by the fact that the $L_1$ error is an integral, and that the integral, in first approximation, is a sum of many independent integrals over small nonoverlapping sets (in the case of a histogram estimate, this statement would be correct if the sample size were Poisson; in the case of the kernel estimate, some extra work is needed). The fact used here is that every data point has only a very local effect. In any case, what matters for now is that for many nonparametric estimates, $E(\int |f_n - f|)$ is indeed a good gauge of the $L_1$ error. Unfortunately, this is not the case for some tailor-designed estimates such as some parametric estimates, but we will nevertheless keep using the mean as our standard of comparison.

The performance of an estimate for a particular $f$ depends upon $f$. It is quite a task to compare estimates with one another, because of this dependence. It is helpful to know how bad **any** estimate has to be as measured by the minimax error

$$\inf_{f_n} \sup_{f \in \mathbf{F}} E(\int |f_n - f|)$$

where $\mathbf{F}$ is a given class of densities. The minimax error, a function of $n$ and $\mathbf{F}$ only, tells us about the error any estimate has to make on at least one density in $\mathbf{F}$. It paints a pessimistic picture, as the $L_1$ error for a given $f_n$ and a given $f \in \mathbf{F}$ can be much smaller than the minimax error. On the other hand, assume that we want to give a person a guarantee (i.e., an upper bound) about the expected $L_1$ error of a given estimate committed on densities in $\mathbf{F}$, and that for that class, the minimax error is 0.23. Then our guaranteed performance cannot be smaller than 0.23. In this respect, minimax errors are often used as lower bounds,

and if minimax errors can't be computed exactly, one should try to compute a lower bound for them.

In **chapter V**, we give a systematic overview of how one can compute such lower bounds by information-theoretic methods. But the main message is contained in the lower bounds themselves. A class $\mathbf{F}$ can be coined a fat class if its minimax error does not tend to zero with $n$. For fat classes, we can't provide any nontrivial performance guarantees for any estimate. What transpires in chapter V is that the class of all densities on [0,1] bounded by 2 and the class of all unimodal densities with infinitely many continuous derivatives are both fat. This implies that to study uniform performances over given classes, imposing tail conditions alone or smoothness conditions alone is not sufficient. At the very least, we need to combine these kinds of conditions. There are even smaller fat classes, such as the class of all normal scale mixtures, the class of all densities whose characteristic function has support on $[-1,1]$, or the class of all densities in an $\epsilon$-ball around a central density $f_0$. The fact that the latter class is fat is not at all surprising since every density is surrounded by millions of ugly densities of very different shapes. A big drawback of the $L_1$ theory is that the $L_1$ distance does not attach a great deal of importance to shape similarities. Shapes can be compared in terms of derivatives. Every $\epsilon$-ball around any density $f_0$ contains many jagged densities $f$ for which the set of points at which $f$ has a derivative has measure zero.

Classes for which the minimax error tends to zero with $n$ include for example classes of densities on [0,1] that are defined via a Lipschitz condition on the $k$-th derivative. Or the class of monotone densities on [0,1] bounded by a constant $B$. Or the class of all normal densities with unknown mean and variance. For all these classes, one may ask for an estimate $f_n$ for which the expected $L_1$ error is uniformly bounded (over $f \in \mathbf{F}$) by a universal constant times the minimax error. Ideally, this constant should be one, but that is often difficult to achieve in practice. Estimates with this property are said to be minimax-optimal for $\mathbf{F}$. They carry a uniform performance guarantee. The construction of such estimates is often ad hoc: sometimes we stumble by accident upon a minimax-optimal estimate (we will show in chapter VII that the kernel estimate is minimax-optimal for many large classes $\mathbf{F}$); sometimes we apply our common sense (this often works for small parametric classes $\mathbf{F}$). But there is also a systematic construction of minimax-optimal estimates, based upon minimum distance estimates, described in **chapter VI**.

The idea is simple enough: cover the set $\mathbf{F}$ by a finite number of $\epsilon$-balls ($\epsilon$ is carefully picked), and define the estimate as the density at one of the centers that is closest to the standard empirical measure according to a criterion that is reminiscent of the $L_1$ criterion. This method requires a finite cover, and hence $\mathbf{F}$ should be $L_1$-totally bounded. The latter restriction can be relaxed, but at some cost.

Chapters V and VI provide us with a lot of information about what can be achieved by density estimates, and what is unreasonable to ask. We give many explicit inequalities, so that one can plug in one's sample size and class descriptors to see what kinds of performance can be expected. If we can make a point

with explicit inequalities, we will do so, for rates of convergence without accurate information about constants and asymptotic error terms are often less appealing to practicing statisticians. Density estimation in general can be compared to an infinite-dimensional parameter-estimation problem. We want to convey to the students in a quantitative fashion, not by experiments, just how difficult density estimation is. How bad are the two troublemakers, the lack of smoothness of $f$, and the size of the tail of $f$? How large should $n$ be for us to be able to do anything meaningful with a given estimate, or any estimate for that matter?

In **chapters VII and VIII**, we study particular estimates. The long chapter VII deals with the kernel estimate. We study the rate of convergence to 0 without trying to exclude any density $f$ from the study. This requires some preliminary work and some generalized definitions of $k$-th derivatives of a density. The effort is well spent. Among other things, we will see that for some kernels, the rate of convergence is limited by the form of the kernel, while for other kernels, the rate is determined solely by the smoothness and the tail of $f$. Always deriving explicit constants, we will see how the shape of $f$ affects the performance, and we can answer such questions as: which density is easiest to estimate by the kernel method for a fixed given kernel $K$?

In **chapter VIII**, we present a case study on monotone density estimation, and compare various estimates for this problem.

Finally, **chapter IX** deals with the issue of relative stability, which was so crucial in the determination of our error criterion (see above).

**Devroye and Gyorfi (1985).**

The $L_1$ approach of the subject is not unlike that of our research monograph with Laszlo Gyorfi ("Nonparametric Density Estimation: The $L_1$ View", John Wiley, 1985, hereafter referred to as **DG**). However, the present text is not a research monograph. We only seek to explain, often sacrificing some deep results for shallower ones with simpler more didactic proofs. The exercises at the end of each chapter should make the text useful for a graduate course on density estimation. We move quickly from one subproblem to another, with very few pauses, trying to maximize the number of ideas and techniques in a book that can be covered in one quarter or trimester.

**DG** is just one of several reference texts that can be considered for further reading and consultation, giving a more comprehensive (but still limited) treatment of the field. It offers a deeper and broader study of some topics touched upon in the present text. For example, **DG** studies more kinds of estimates, such as automatic kernel estimates (**DG**, chapter VI), generalized kernel and histogram estimates (**DG**, chapter VII), transformed kernel estimates (**DG**, chapter IX) and orthogonal series estimates (**DG**, chapter XII). In addition, **DG** discusses several applications, including simulation (**DG**, chapter VIII), discrimination (**DG**, chapter X) and detection (**DG**, chapter XI). The present text is organized

like lecture notes for graduate students. The readers will however appreciate that we have kept the notation consistent throughout both books.

There is some new material presented here, not found in **DG**, such as the introductory notions on robustness (chapter IV) and minimum distance estimation (chapter VI), and the in-depth studies of monotone density estimation (chapter VIII) and relative stability (chapter IX).

The notes are geared towards students who have never been exposed to density estimation, but who do have a basic background in analysis and probability.

## Applications.

**Exploratory data analysis** is concerned with the extraction of information from data in order to choose appropriate statistical procedures for analyzing the data. Obviously, nonparametric density estimates seem prime candidates for such a quick analysis, especially when combined with a good graphics package and a friendly workstation.

We should not forget the important role of densities and their estimates in **probability theory**. Density estimates define smoothed empirical measures. For the ordinary atomic empirical measure $\mu_n$ (which puts mass $1/n$ at each data point), we have the disappointing property that

$$\sup_A \; | \mu_n(A) - \mu(A) | \; = 1 \; ,$$

where $\mu$ is the measure induced by a density $f$, and the supremum is over all Borel sets $A$. However, if the expected $L_1$ error for $f_n$ tends to 0, then

$$\lim_{n \to \infty} E\left( \sup_A \; \left| \int_A f_n - \int_A f \; \right| \right) = 0 \; .$$

Thus, when densities exist, $f_n$ defines an **empirical measure** that is more precise than the standard empirical measure. Interestingly, this property does not necessarily carry over when the collection of Borel sets in the supremum is replaced with all left-infinite intervals $(-\infty, x]$, yielding the Kolmogorov-Smirnov statistic.

If one has drawn samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ from unknown densities $f$ and $g$ respectively, and one has to decide which density ($f$ or $g$) a new sample $Z_1, \ldots, Z_k$ is drawn from, the maximum likelihood principle ("choose $f$ if $\prod f(Z_i)/g(Z_i) \geq 1$ ") is not directly applicable. Replacing $f$ and $g$ by estimates $f_n$ and $g_n$ in the likelihood products is risky. This issue and some solutions to the **detection problem** are discussed in chapter XI of **DG**.

Perhaps the most popular application of density estimation is in **discrimination** (or: **pattern recognition**). Assume that $X_i$ and $Y_i$ samples are given as in the detection problem, and that one is presented with a single random variable $Z$, which is known to have density $f$ or $g$. The best possible (or: Bayes)

rule for this problem states that we should assign $Z$ to $f$ if $f(Z) \geq g(Z)$, and to $g$ otherwise. It should be obvious that if $f$ and $g$ are unknown, then suitable estimates of them can lead to rules that perform almost as good as the Bayes rule. Both the detection and discrimination problems require derivatives of probability measures, and cannot be solved without directly or indirectly estimating or approximating these derivatives. Good starting points for further study are Cover (1969), Cover and Wagner (1975) and Duda and Hart (1973).

**Estimating a tail probability** based upon a sample of size $n$ drawn from an unknown density $f$ can be done by a frequency count: $\int_x^\infty f$ is estimated by $1-F_n(x)$ where $F_n$ is the standard empirical distribution function. When tail probabilities are needed for testing it becomes important to have estimates that are smooth and have small relative errors and variances. In such cases, one can replace $1-F_n(x)$ by $\int_x^\infty f_n$ where $f_n$ is an appropriate density estimate, which is sometimes based upon a model one has constructed of the shape and size of the tail of $f$.

**Clustering** is one of the main tools of data analysis. Data points with an unknown density $f$ can be clustered by an analysis of the estimated shape of $f$. Good candidates for cluster centers are the modes (or peaks) of $f$; hence the need to estimate modes. Separating boundaries between clusters roughly correspond to the valleys of $f$; hence the need to reconstruct the valleys of a density. Mode estimation is discussed by Wegman (1972), clustering is studied by Hartigan (1975), and the connection with density estimation is apparent from Hartigan (1981).

Some applications demand the evaluation of a **functional** of $f$, such as $\int f^2$, $\int (f'^2/f)$ (Fisher's information criterion), or $\int f \log f$ (minus the entropy of $f$). These quantities are important in classical statistics. For example, Fisher's information criterion appears in the Cramer-Rao inequality (see e.g. Pitman, 1979). The entropy is related to the performance of maximum likelihood estimates. The estimation of these functionals from a sample can help statisticians in their decision-making processes.

In **simulation**, the following problem often occurs: data $X_1, \ldots, X_n$ are collected at some cost, and it is assumed that the $X_i$'s are iid with a common but unknown density $f$. In a simulation run, new data from $f$ are required, but it is too expensive or unfeasible to collect new data in the field. Rather, one is forced to do with the available information. If $k$ new data points are generated from a density estimate $f_n$, then one commits an error. This error can be measured in terms of the minimal number of the newly generated data that need to be replaced by other points in order to turn the new sample (from $f_n$) into a sample from $f$. It turns out that the minimal number is binomially distributed with parameters $k$ and $\frac{1}{2}\int |f_n - f|$, which once again points to the importance of the $L_1$ criterion. See chapter VIII of **DG**.

Estimating the **shape** of a density is much more difficult than estimating a density in $L_1$, for shapes are alike if all (or most) derivatives are alike, assuming that derivatives exist. Thus, in shape estimation, distances between densities should be measured in terms of derivatives as is done for example in Sobolev spaces (Adams, 1975). As an alternative, we could compare the frequency spectra of two densities (to draw an analog, consider that close frequency spectra of voice signals indicate that the voice signals have similar shapes). But the frequency spectrum of a density is nothing but its characteristic function $\phi$. If our estimate $f_n$ has characteristic function $\phi_n$, then the distance between $f_n$ and $f$ can be defined by

$$\sup_t \; | \phi_n(t) - \phi(t) | \; .$$

**Further work.**

All the applications mentioned above require density estimates, but each application imposes its individual demands on the estimate. Only a few of them require densities that are close in the $L_1$ sense. Hence the need to study density estimates that are good in other respects. Because of the crucial and natural role played by $L_1$, it is felt that the results from the $L_1$ theory will aid substantially in the derivation and understanding of non-$L_1$ properties.

Many estimates are tailor-designed for target classes $\mathbf{F}$. They are often useless outside the target classes. On the other hand, nonparametric estimates are reliable for all $f$ but generally speaking inferior on $\mathbf{F}$ to a tailor-designed estimate for that class. Thus, we should try to combine estimates in such a way that on $\mathbf{F}$, the resulting estimate performs as the tailor-designed estimate, and outside $\mathbf{F}$, it inherits the consistency and rate of convergence of the nonparametric estimate. Such combined estimates are for example required in automatic computer packages for density estimation. The choice of an estimate can be based upon the $L_1$ distance between the various estimates: nonparametric estimates could be replaced by tailor-designed estimates if they fall in an $\epsilon$-ball (a halo) around one of the latter estimates, where $\epsilon$ is a carefully picked radius of influence. See Devroye (1986).

Some researchers take the point of view that we should first choose a class of estimates, and then try to make the best of it. They are willing to accept the consequences of this strategy, i.e. their expected $L_1$ error is bounded from below by

$$V(f,n) \stackrel{\Delta}{=} \inf_{f_n \in \mathbf{C}} E(\int | f_n - f \; |) \, ,$$

where $\mathbf{C}$ is the class of estimates under consideration. For some classes $\mathbf{C}$, $V(f,n)$ can be readily computed. Sometimes it is even possible to compute non-trivial lower bounds for $V(n) \stackrel{\Delta}{=} \inf_f V(f,n)$. $V(n)$ indicates the absolute

limitations of the given class of estimates. Even if one were shown which $f$ is being estimated, and picked the best $f_n$ in $\mathbf{C}$ accordingly, the expected $L_1$ error would still have to be at least $V(n)$. It is instructive to compute $V(n)$ for many popular classes. For example, for the class $\mathbf{C}$ of all kernel estimates,

$$V(n) \geq \frac{1}{\sqrt{528\ n}}$$

(Devroye, 1986). If only nonnegative kernels are allowed, we have for all $f$,

$$V(f,n) \geq \frac{0.86+o(1)}{n^{\frac{2}{5}}}$$

(Devroye and Penrod, 1984). The last inequality basically implies that to make the expected $L_1$ error less than 0.01 with such kernel estimates, $n$ should be at least 100,000, regardless of how nice the density is that is being estimated. One can consider these lower bounds as the costs associated with the use of the kernel estimate (nothing is free). It is also necessary to verify what costs are associated with the estimation of one particular $f$, as measured by $V(f,n)$. For the kernel estimate, an $L_2$ analog of $V(f,n)$ has been computed in a milestone paper by Watson and Leadbetter (1963) (see also the follow-up papers by Davis (1975,1977)), but $V(f,n)$ seems much more difficult to compute in $L_1$. A detailed study of $V(f,n)$ is essential for a solid understanding of the kernel estimate.

The sample size restrictions imposed by the kernel estimate even on the best $f$ are nearly unacceptable. We feel therefore that the major practical breakthroughs in density estimation will not be on nonparametric estimates in their general forms, but on tailor-designed estimates. The target classes on which advances should be made are somewhere in the grey area between finite-parameter classes and fat classes. A prime example of this is the class of monotone densities, for which good estimates were developed by Grenander (1956) and Birge (1983,1984). Particular versions of the kernel estimate perform well for target classes defined in terms of the smoothness of $f$, as is shown in chapter VII. However, how does one handle important classes such as the class of all log-concave densities (this class includes normal, gamma, Weibull, beta, exponential power, logistic, hyperbolic secant, generalized inverse gaussian, extreme value and Perks distributions), the class of all normal scale mixtures, or the class of all densities of sums of $k$ iid random variables with density supported on [0,1]?

With every class $\mathbf{C}$ of estimates we have associated lower performance bounds $V(f,n)$ and $V(n)$. It would be nice if, based upon the data, we could select an $f_n$ from $\mathbf{C}$ for which the lower bound $V(f,n)$ is attained modulo a multiplicative constant. Such a data-based selection rule for $\mathbf{C}$ is exciting since we would in fact be able to obtain the best possible rate of convergence within $\mathbf{C}$, sometimes without knowing what the rate of convergence is. A good starting point for further research is Stone (1984,1985) where a similar problem is successfully solved in $L_2$ for the kernel and histogram estimates.

Let us conclude by mentioning the difficult problem of the estimation of $\int |f_n - f|$, on which absolutely no headway has been made to date. It should be noted that if we could estimate $\int |f_n - f|$ accurately (e.g., with an expected error much smaller than $E(\int |f_n - f|)$), then we would be able to make a lot of progress in the data-based selection problem mentioned in the previous paragraph.

**Dedication.**

This book is dedicated to Bea.

Luc Devroye
School of Computer Science
McGill University
Montreal, Canada H3A 2K6

# Chapter One
# DISTANCES BETWEEN DENSITIES

The distance between two densities $f$ and $g$ on the Borel sets of $R^d$ can be defined in many different ways. First and foremost, there is the $L_p$ distance

$$L_p(f,g) = \begin{cases} \left(\int |f-g|^p\right)^{1/p} & (\infty > p > 0) \\ \text{ess sup } |f-g| & (p=\infty) \end{cases}$$

The Hellinger distance is defined by

$$H_p(f,g) = \left(\int (f^{1/p} - g^{1/p})^p\right)^{1/p} \quad (p>0).$$

Finally, the distance between $f$ and $g$ can also be measured in terms of an entropy-related quantity, the **Kullback-Leibler** number

$$K(f,g) = \begin{cases} \int f \log \frac{f}{g} & \text{if } f << g \\ \infty & \text{otherwise} \end{cases}.$$

In this introductory chapter, we will establish relationships between these quantities, and explore the properties of $L_1$ in more detail.

## 1.1. THE TOTAL VARIATION.

The **total variation** between two probability measures $\mu$ and $\nu$ on the Borel sets of $R^d$ is defined by

$$V = \sup_A |\mu(A) - \nu(A)|.$$

It should be noted that $0 \leq V \leq 1$ in all cases, and that it is well-defined even if $\mu$ and/or $\nu$ do not have densities. When $V$ is 0.01, we know that for any set $A$,

the probability assigned to it by $\mu$ differs at most by 0.01 from the probability assigned to it by $\nu$. In other words, $V$ is a practical easy-to-understand quantity. When $\mu$ and $\nu$ have densities $f$ and $g$, we have

**Theorem 1.1. Scheffe's theorem.**

$$\int |f-g| = 2 \sup_A \left| \int_A f - \int_A g \right|$$
$$= 2\int (f-g)_+ = 2\int (g-f)_+ .$$

**Proof of Theorem 1.1.**

Intuitively, Scheffe's theorem is clear from figure 1.1.

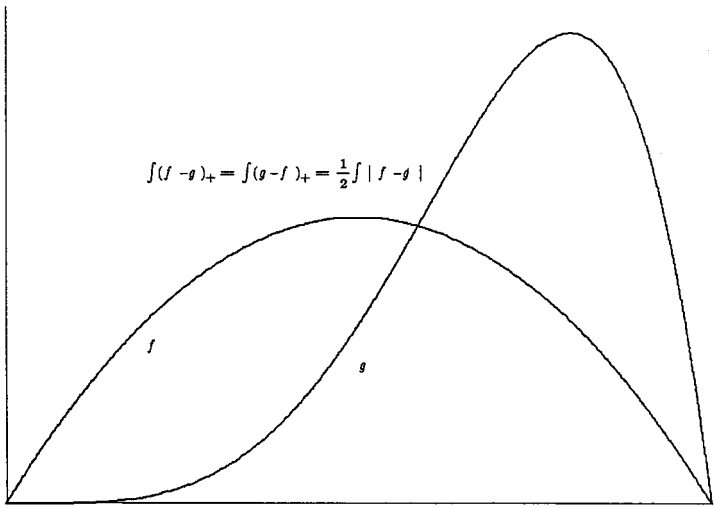


**Figure 1.1.**
Two densities.

The integrals of $(f-g)_+$ and $(g-f)_+$ are equal, and sum to $(1/2)\int |f-g|$. Scheffe's theorem states that the supremum is reached for the set $B=\{f>g\}$ (or, equivalently, for the set $\{g>f\}$). To prove this formally, consider first that

$$\int |f-g| = 2\int_B (f-g) \leq 2 \sup_A \left| \int_A f - \int_A g \right| .$$

To prove the other inequality, note that

$$\begin{aligned} \left| \int_A f - \int_A g \right| &= \left| \int_{A\cap B} (f-g) + \int_{A\cap B^c} (f-g) \right| \\ &\le \max\left( \int_{A\cap B} (f-g), \int_{A\cap B^c} (g-f) \right) \\ &\le \max\left( \int_B (f-g), \int_{B^c} (g-f) \right) \\ &= \frac{1}{2}\int |f-g| \,. \blacksquare \end{aligned}$$

One of the most remarkable properties of $L_1$ is that $L_1(f,g)$ is invariant under rich transformations of the coordinate axes, where a transformation $T$ is called **rich** if $\{T^{-1}B \mid B\in\mathbf{B}\}=\mathbf{B}$, and $\mathbf{B}$ denotes the Borel sets of $R^d$. Note that this necessarily implies that the mapping is one-to-one. For example, a transformation in which each coordinate is transformed separately via a strictly increasing mapping will do. Assume that random variables $X,Y$ have densities $f,g$, and that $T(X),T(Y)$ have total variation $V$. Then

$$\begin{aligned} \int |f-g| &= 2\sup_A |P(X\in A)-P(Y\in A)| \quad \text{(Scheffe's theorem)} \\ &= 2\sup_A |P(T(X)\in T(A))-P(T(Y)\in T(A))| \quad \text{(one-to-one mapping)} \\ &= 2\sup_A |P(T(X)\in A)-P(T(Y)\in A)| \quad (\{T(A)\}=\{A\cap T(R^d)\}) \\ &= 2V \,. \end{aligned}$$

This means that the $L_1$ distance between $f$ and $g$, hidden in infinite tails, can be visualized by plotting the densities of transformed random variables. For example, the transformation $x:\to x/(1+|x|)$ maps the real line to $[-1,1]$, and the tails show up at near the ends of this interval. This is especially useful for displaying infinite-tailed densities on a terminal screen.

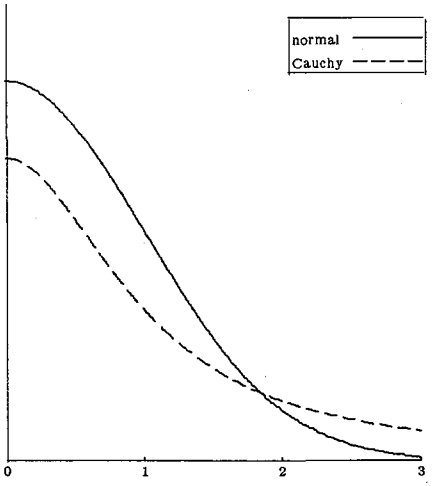


**Figure 1.2a.**
Two densities on the real line.

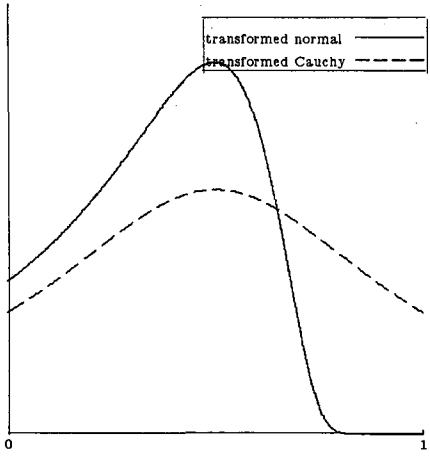


**Figure 1.2b.**
Densities of figure 1.2a after transformation $y := x/(1+|x|)$.

## 1.2. THE SPACE $L_p$.

Needless to say, $L_1$ is the natural space for all densities. When we mention something involving $L_p(f,g)$ for some $p \neq 1$, it is always understood that $f, g \in L_p$, i.e. $\int f^p < \infty, \int g^p < \infty$.

Let us first note that there cannot exist any direct inequalities between $L_1(f,g)$ and $L_p(f,g)$ for $p \neq 1$. To see this, it suffices to note (see figure 1.3 below)

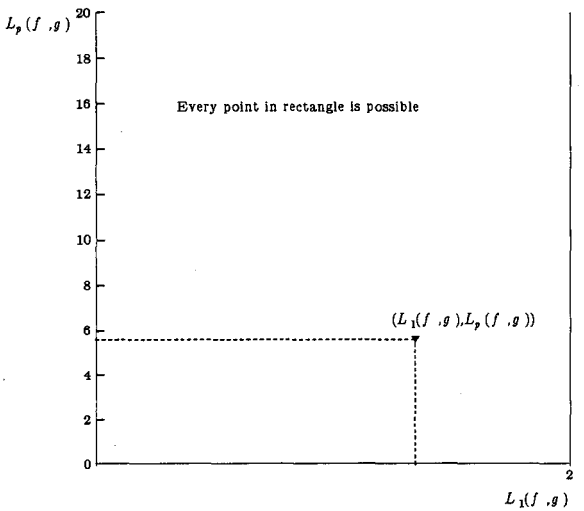


**Figure 1.3.**
$(L_1, L_p)$ plane

that for every point in the plane $[0,2]\times[0,\infty)$, there exists a pair of densities $(f,g)$ such that the $x$-coordinate is $L_1(f,g)$, and the $y$-coordinate is $L_p(f,g)$. Once a pair $(f,g)$ is fixed, changing the scale takes the point along a vertical journey in the plane. This follows from the observation that if $f,g$ are the densities of $X,Y$, and $f^*,g^*$ are the densities of $aX,aY$ where $a$ is a real number, then

$$L_p(f^*,g^*) = a^{\frac{1-p}{p}} L_p(f,g).$$

In other words, only $L_1$ is unaffected by a rescaling. And thus, only $L_1(f,g)$ can have a universal interpretation.

If one density (say, $f$) is kept fixed, we can still cover most of the plane by varying $g$, although there are some restrictions. These are perhaps best captured in

**Theorem 1.2.**

Let $f$ be a fixed density on $R^d$, and let $p > 1$. Assume that $f \in L_p$. Then there exists a sequence of densities $f_n \in L_p$ such that

$$L_1(f_n, f) \to 0, \quad L_p(f_n, f) \to \infty.$$

However, the converse is not true: if $L_p(f_n, f) \to 0$ as $n \to \infty$, then $L_1(f_n, f) \to 0$.

**Proof of Theorem 1.2.**

Find a sequence of densities $g_n$ for which $(\int g_n{}^p)^{1/p} = a_n \to \infty$. Define $b_n = \min(1, 1/\sqrt{a_n})$. Set

$$f_n = f(1-b_n) + g_n b_n.$$

We verify easily that

$$\int |f_n - f| \le 2b_n \to 0,$$

$$\left(\int |f_n - f|^p\right)^p = \left(\int |b_n g_n - b_n f|^p\right)^p$$

$$\ge b_n \left((\int g_n{}^p)^{1/p} - \left((\int f^p)^{1/p}\right) \to \infty \quad \text{(Minkowski)}.$$

This proves the first part of the theorem. For the second part, we take a number $M$ so large that $2\int_{[-M,M]} f$ is as small as desired. Then,

$$\int |f_n - f| = 2\int (f - f_n)_+ \le 2\int_{[-M,M]} + 2\int_{[-M,M]^c}$$

$$\le 2\left(\int_{[-M,M]} dx\right)^{\frac{1}{q}} \left(\int |f_n - f|^p\right)^{\frac{1}{p}} + 2\int_{[-M,M]^c} f \quad \left(\text{Holder}; \frac{1}{p} + \frac{1}{q} = 1\right)$$

$$= o(1) + 2\int_{[-M,M]^c} f.$$

This concludes the proof of Theorem 1.2. ■

## 1.3. HELLINGER SPACES.

$H_p$ shares many nice features with $L_1$: for any pair $(f, g)$, $H_p(f, g)$ is non-negative and finite, the maximum possible value being $2^{1/p}$. It is easy to verify that $H_p$ remains invariant under "rich" transformations, and is thus scale-invariant. This can be proved based upon a relationship between $H_p$ and the

probabilities of sets. See exercise 1.2, for example. The most important values for $p$ are 1 and 2. Clearly, $H_1(f,g)=L_1(f,g)$. Furthermore,

**Theorem 1.3.**

For any pair of densities $(f,g)$,

$$H_2^2(f,g) \le H_1(f,g) \le H_2(f,g)\sqrt{4-H_2^2(f,g)} \le 2H_2(f,g)$$

and

$$2-H_1(f,g) \ge \left(1-\frac{1}{2}H_2^2(f,g)\right)^2 .$$

The last inequality is equivalent to **LeCam's inequality** (LeCam, 1973)

$$\int \min(f,g) \ge \frac{1}{2}\left(\int\sqrt{fg}\right)^2 .$$

**Proof of Theorem 1.3.**

In the proof, we will drop the arguments $(f,g)$. We have

$$\begin{aligned} H_1 &= \int |f-g| = \int |\sqrt{f}-\sqrt{g}|\,|\sqrt{f}+\sqrt{g}| \\ &\ge \int(\sqrt{f}-\sqrt{g})^2 = H_2^2 . \end{aligned}$$

Also,

$$\begin{aligned} H_1^2 &\le \int(\sqrt{f}-\sqrt{g})^2 \int(\sqrt{f}+\sqrt{g})^2 \quad \text{(Cauchy-Schwarz)} \\ &= H_2^2(2+2\int\sqrt{fg}) = H_2^2(4-H_2^2) . \end{aligned}$$

LeCam's inequality can be obtained very simply as follows:

$$\left(\int_{f<g}\sqrt{fg}\right)^2 = \left(\int_{f<g} f\sqrt{\frac{g}{f}}\right)^2$$

$$\le \int_{f<g} f\frac{g}{f} \int_{f<g} f \le \int_{f<g} f \quad \text{(Cauchy-Schwarz)} .$$

Combining this with a similar inequality for the set $\{f \ge g\}$, we have

$$\left(\int\sqrt{fg}\right)^2 \le 2\int_{f<g} f + 2\int_{g\le f} g = 2\int\min(f,g) .$$

Finally, use the fact that $\int\min(f,g)=1-H_1/2$, and that $H_2^2=2-2\int\sqrt{fg}$. ■

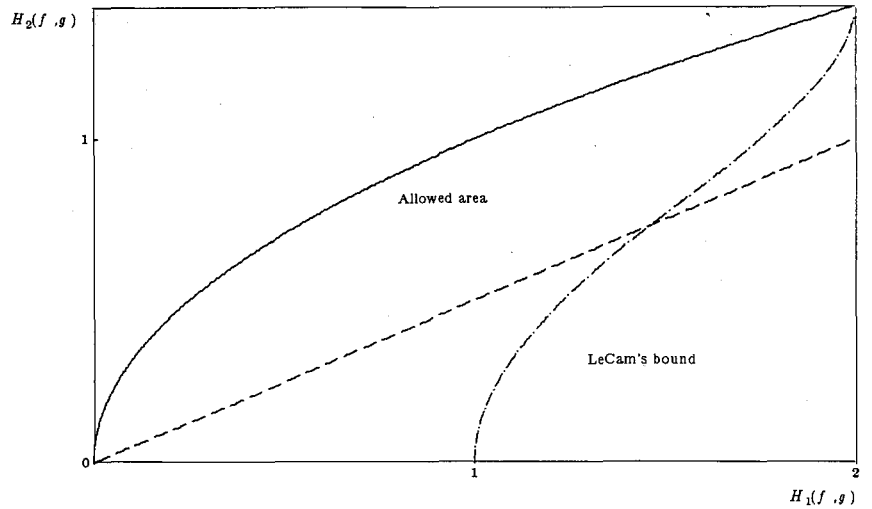


**Figure 1.4.**
$(H_1, H_2)$ plane

Let us interpret these inequalities in the plane formed by $H_1$ and $H_2$, i.e. in $[0,2]\times[0,\sqrt{2}]$. Figure 1.4 above shows that only a thin banana-shaped area around the diagonal is possible. In particular, $H_2$ and $H_1$ have to converge to zero together. This could be a potentially useful fact for proving the consistency in one metric using consistency in the other metric. A word of caution. The rates of convergence to zero can differ widely, since $H_1$ can be made to decrease linearly or quadratically in $H_2$. LeCam's inequality cuts out an area near the top right corner of the rectangle, providing us with information regarding the rate with which $2-H_1$ tends to zero when $\sqrt{2}-H_2$ tends to zero. At this point, the reader may wonder why one would be interested at all in that part of the plane, $H_1$ being so far away from zero. Observe that if $f\ ,g$ are replaced by $n$-fold products (as in the case of a sample of $n$ iid random variables drawn from $f$ or $g$), then the $H_p$-values for the products are pushed towards their upper bounds as $n$ increases (unless $f=g$). It is precisely in those situations that we will need the behavior of $H_p$ near $2^{1/p}$.

## 1.4. ENTROPY. KULLBACK-LEIBLER NUMBERS.

The quantity $\int f \log f$ has often been related to the **entropy** of a density $f$ (actually, minus this quantity is the standard definition). Note that the function $x \log x$ for $x \geq 0$ is convex, taking the value 0 at 0, dipping underneath the axis, reaching a minimum $-1/e$ at $x = 1/e$, crossing the axis again at $x = 1$, and increasing to $\infty$ as $x \to \infty$. Thus, there are no problems with $0 \log 0$ in the integral, since this should clearly be interpreted as 0.

$\int f \log f$ can take all values in $[-\infty, \infty]$. To force it to take the value $\infty$, an infinite peak is needed. To force it to take the value $-\infty$, an infinite tail is required, i.e. the probability mass has to be smeared out thinly. Note that there could be a possible problem with the integral when both its positive and negative components are $\infty$.

The **Kullback-Leibler** number

$$K(f, g) = \int f \log \frac{f}{g},$$

defined when $f \ll g$ ($f$ is absolutely continuous with respect to $g$) is asymmetric in $f, g$, and can therefore not be a distance. Nevertheless, there are important relationships between $K$ and $L_1$. Its main use is when $f, g$ are replaced by $n$-fold products. The Kullback-Leibler number for the product densities is precisely $n$ times $K(f, g)$. This facilitates certain computations dramatically.

As with all integrals that can be written as $\int f \psi(\frac{f}{g})$ for some function $\psi$ (examples include $K$ and $H_p$ for all $p$), $K$ is invariant under rich transformations. In other words, $K(f, g)$ depends on the relative shapes of $f, g$ only. Note first that $K \geq 0$, with equality occurring if and only if $f = g$. This follows from Jensen's inequality:

$$-K(f, g) = \int f \log \frac{g}{f} \leq \log(\int f \frac{g}{f}) = 0.$$

Note however that $K(f, g)$ can be $\infty$ even though $f \ll g$. This means that the $(L_1, K)$ plane cannot possibly be restricted from the top in the $y$-direction. It is true however that when $K$ is small, $L_1$ has to be small. Several inequalities can be invoked to decide just how small $L_1$ has to be.

**Theorem 1.4.**

Let $K = K(f, g)$, and $L_1 = L_1(f, g)$. Then

$$L_1 \leq \sqrt{2K}$$

(the Kullback-Csiszar-Kemperman inequality; see Kullback (1967), Csiszar (1967), Kemperman (1969)), and

$$L_1 \leq 2\sqrt{1-e^{-K}} \leq 2-e^{-K} .$$

(the Bretagnolle-Huber inequalities; see Bretagnolle and Huber, 1979). The last inequality can be restated as follows: $\int \min(f, g) \geq \frac{1}{2} e^{-K(f,g)}$.

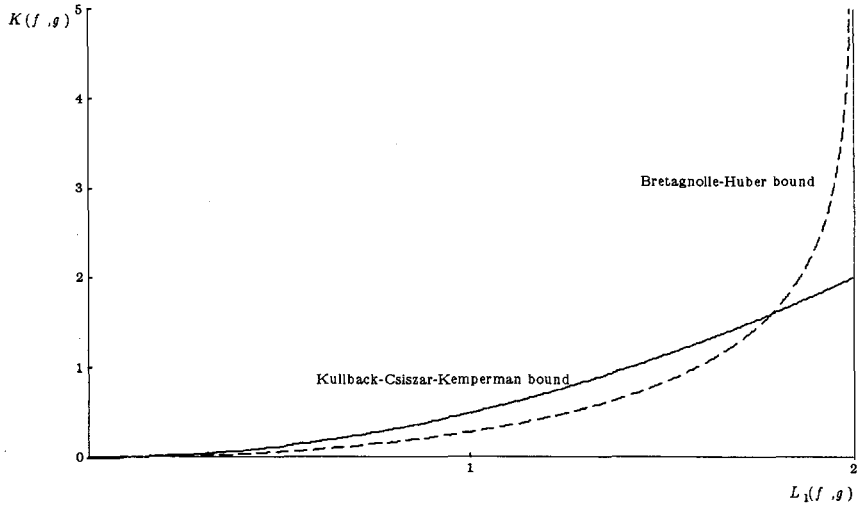


**Figure 1.5.**
$(L_1, K)$ plane

**Proof of Theorem 1.4.**

We begin with the Bretagnolle-Huber inequalities. By a simple application of Jensen's inequality, the following is true:

$$-\int f \log\frac{f}{g} = \int f \log(\min(\frac{g}{f},1))+\int f \log(\max(\frac{g}{f},1))$$

$$\leq \log(\int \min(f,g))+\log(\int \min(f,g)) = \log((1-\int |f-g|/2)(1+\int |f-g|/2))$$

$$= \log(1-\frac{1}{4}(\int |f-g|)^2) .$$

The second inequality of Bretagnolle and Huber is rather obvious. Turning to the Kullback-Csiszar-Kemperman inequality, we introduce the following notation: $I$ is the indicator function, $A=\{f \geq g\}$, $h=gI_A/q$, $q=\int_A g$, $p=\int_A f$. First,

$$\int_A f \log\frac{f}{g} = \int h\frac{f}{g}\log\frac{f}{g}\int_A g$$

$$\geq \int h\frac{f}{g}\log(\int h\frac{f}{g})\int_A g \quad \text{(Jensen's inequality)}$$

$$= \int_A f \log(\int_A f / \int_A g) = p\log\frac{p}{q} .$$

Thus, repeating this argument for the complement of A, we have

$$K(f,g) \geq p\log\frac{p}{q}+(1-p)\log\frac{1-p}{1-q} = H(p,q) \quad \text{(definition of } H(p,q)) .$$

Assume that $p=q+r$ for some $r>0$. Then, writing $H(r)$ instead of $H(p,q)$, we have

$$H(r) = (q+r)\log(1+\frac{r}{q})+(1-q-r)\log(1-\frac{r}{1-q}) ,$$

$$H'(r) = \log(1+\frac{r}{q})-\log(1-\frac{r}{1-q}) ,$$

$$H''(r) = \frac{1}{p(1-p)} \geq 4 .$$

Thus, by Taylor's series expansion with remainder,

$$K(f,g) \geq 4\frac{r^2}{2} = 2(p-q)^2 = \frac{1}{2}\left(\int |f-g|\right)^2 . \blacksquare$$

Figure 1.5 shows the areas cut out by the inequalities of Theorem 1.4.

## 1.5. STANDARD IMPROVEMENTS OF DENSITY ESTIMATES.

The question dealt with in this section is very simple: if $g$ is replaced by a common sense improvement, is the new improved $g$ closer to $f$ than the old $g$? Very often, the answer is positive.

For example, it cannot hurt to restrict a density estimate to the support of $f$.

**Theorem 1.5.**

Let $g$ be a function with $\int g = 1$, and let $f$ be a density vanishing off $S$. Define

$$g^* = gI_S / \int_S g .$$

Then $\int | g^* - f | \leq \int | g - f |$.

**Proof of Theorem 1.5.**

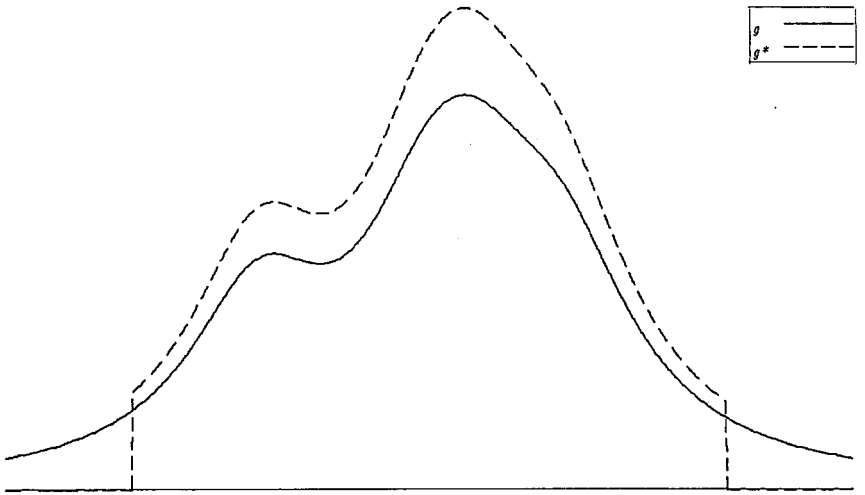


**Figure 1.6.**
Density restricted to support of another density.

$$\int | f - g^* | = 2\int_S (f - \frac{g}{\int_S g})_+ \quad \text{(Scheffe's theorem)}$$

$$\leq 2\int_S (f-g)_+ \leq 2\int (f-g)_+ = \int |f-g| \,. \blacksquare$$

It also can't hurt to normalize a density.

**Theorem 1.6.**

Let $g$ be a function with $\int g = 1$, and let $f$ be a fixed density. Define

$$g^* = gI_S / \int_S g$$

where $S = \{g > 0\}$. Then $\int |g^* - f| \leq \int |g - f|$.

**Proof of Theorem 1.6.**

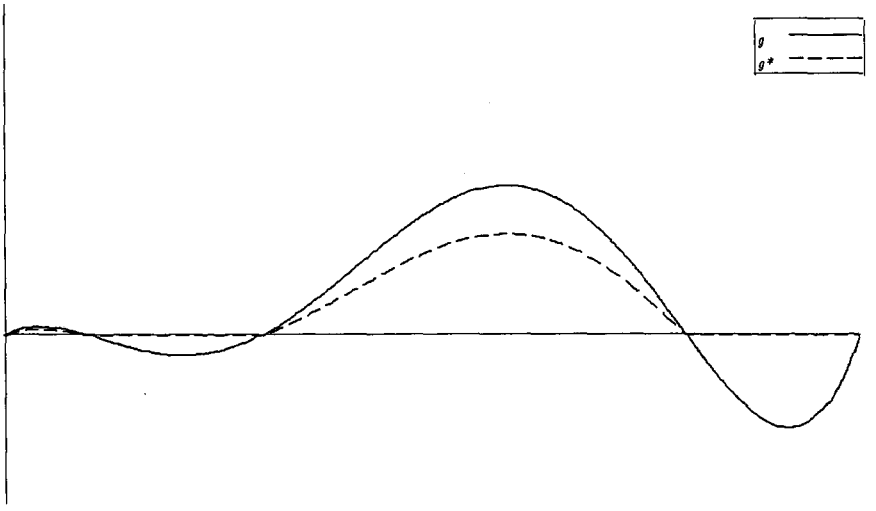


**Figure 1.7.**
Function integrating to one, and corresponding normalized density.

$$\int |f - g^*| = 2\int_S (g^* - f)_+ \quad (\text{use } \int g = 1)$$

$$= 2\int_S (g^* - f)_+ + 2\int_{S^c} (g^* - f)_+$$

$$= 2\int_S (g^* - f)_+$$

$$\leq 2\int_S (g - f)_+ = 2\int (g - f)_+ = \int |f - g| \,. \blacksquare$$

For other possible reasonable improvements, see exercises 1.5 and 1.6.

## 1.6. PROJECTION ESTIMATES.

Assume that $f_n$ is a density estimate, and that $\mathbf{F}$ is a class of densities containing $f$. If we know $\mathbf{F}$, then we would often like to estimate $f$ by another member of $\mathbf{F}$. Think for example about the class of all unimodal densities on the real line. The **projection** of $f_n$ onto $\mathbf{F}$ is $g_n$ where

$$\int |g_n - f_n| \stackrel{\Delta}{=} \inf_{g \in \mathbf{F}} \int |g - f_n| \,.$$

For very small (usually parametric) classes, the projection $g_n$ is usually much closer to $f$ than $f_n$. This has been exploited by Beran (1977, 1981) for obtaining robust parametric estimates. When $\mathbf{F}$ is fairly rich, it turns out that $g_n$ inherits the rate of convergence of $f_n$. In other words, projection estimates are "safe" if we know that $f \in \mathbf{F}$. The statement about the rate of convergence is implicit in

**Theorem 1.7.**

For all $f \in \mathbf{F}$, the projection $g$ of $h$ onto $\mathbf{F}$ satisfies:

$$\int |g - f| \leq 2 \int |h - f| \,.$$

**Proof of Theorem 1.7.**

$$\int |g-f| \leq \int |h-f| + \int |h-g|$$
$$\leq 2 \int |h-f| . \blacksquare$$

## 1.7. EXERCISES.

1.1. Construct a density which does not belong to any $L_p$ for $p \neq 1$.

1.2. Show that

$$H_2^2(f,g) = 2 \sup_{\text{all partitions } A_1, \ldots, A_n \text{ of } R^d} \left(1 - \sum_{i=1}^{n} \sqrt{\int_{A_i} f \int_{A_i} g}\right).$$

See figure 1.8.

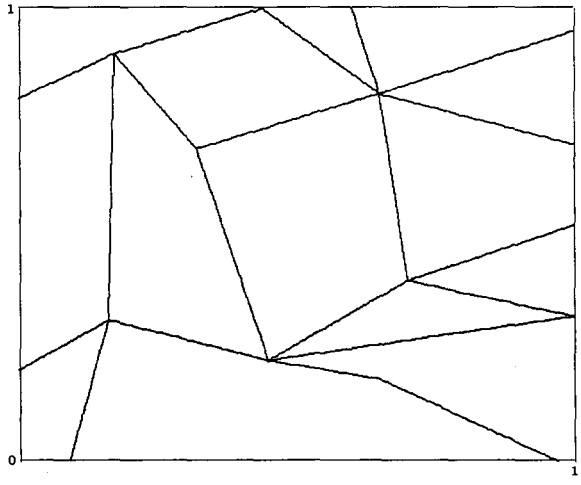


**Figure 1.8.**
Partition of $[0,1] \times [0,1]$

Note that this equality can be used to define the $H_2$ distance between two measures even if densities do not exist.

1.3. Let $T$ be any mapping: $R \to R$. Let $X, Y$ have densities $f, g$, and let $T(X), T(Y)$ have probability measures $\mu, \nu$. Using the total variation interpretation of $H_1$, and the property of $H_2$ given in exercise 1.2, show that

$$H_p(\mu,\nu) \leq H_p(f,g)$$

for $p = 1,2$.

1.4. Prove that $K(f,g) \geq -2\log(1-H_2^2(f,g)/2) \geq H_2^2(f,g)$.

1.5. Prove or disprove: assume that $g \geq 0, \int g < 1$. Let $g^*$ be any density satisfying $g^* \geq g$. Then $\int |g^*-f| \leq \int |g-f|$ for any density $f$.

1.6. Let $f \geq 0$ be $\downarrow$ on $[0,1]$, and let $g \geq 0$ be $\uparrow$ on $[0,1]$. Define $g^*$ as the function with constant value $\int g$ on $[0,1]$ and the value 0 outside $[0,1]$. Show that

$$\int |g^*-f| \leq \int |g-f| .$$

# Chapter Two
# DENSITY ESTIMATION AND DERIVATION OF MEASURES

## 2.1. OUR MODEL.

We assume that a sample $X_1, \ldots, X_n$ of iid random vectors with unknown density $f$ is given. Unless explicitly mentioned, this is the only information available to us regarding $f$. A **density estimate** $f_n$ is a Borel measurable function of $x$ and the **data** $X_1, \ldots, X_n$:

$$f_n = f_n(x, X_1, \ldots, X_n).$$

Usually, $f_n$ is a density in $x$, i.e. it is nonnegative and integrates to one.

When $f$ has probability measure $\mu$, and $\lambda$ is Lebesgue measure, then $f$ is almost everywhere equal to the Radon-Nikodym derivative $\frac{d\mu}{d\lambda}$. This means that for all Borel sets $A$,

$$\int_A f = \mu(A).$$

We will see that most estimates attempt to approximate the derivative of $\mu$ with respect to $\lambda$.

## 2.2. POPULAR DENSITY ESTIMATES.

The **empirical measure** $\mu_n$ is the frequency count of a set, divided by $n$:

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} I_A(X_i).$$

It can be used as an approximation of $\mu(A)$. Thus, if we partition the space $R^d$ into a countable collection $A_1, A_2, \ldots$, and all $A_i$'s are small, then

$$f_n(x) = \frac{\mu_n(A_i)}{\lambda(A_i)}, \quad x \in A_i,$$

can be expected to be a reasonable approximation of $\frac{d\mu}{d\lambda}$. For the approximation to be good, we need two things:

A. Every $\mu(A_i)/\lambda(A_i)$ should be close to $f$, i.e. the $A_i$'s should be small.

B. $\mu_n(A_i)/\mu(A_i)$ should be close to its mean (1). For this to be true, $n$ should be large.

The estimate $f_n$ defined by a frequency count on a partition is called the **histogram estimate.** An example with $d=1$ is shown in figure 2.1.

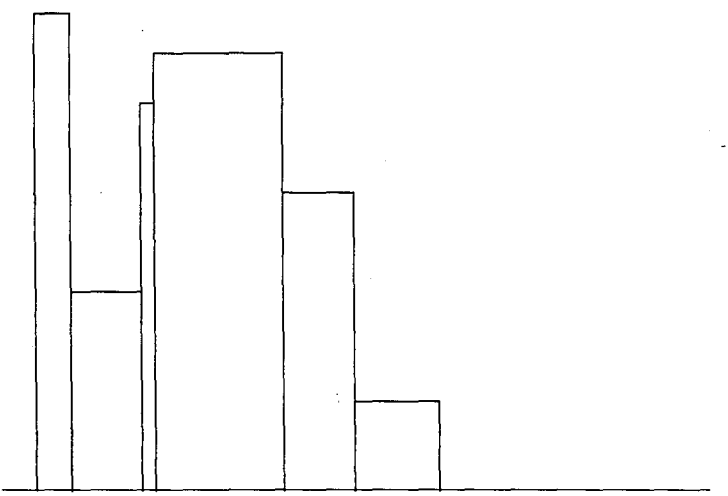

**Figure 2.1.**
Histogram estimate.

Typically, the partition consists of a regular grid of equal size intervals or hyperrectangles. We will encounter situations in which it is to our advantage not to choose intervals of the same size. Furthermore, the partition itself can sometimes depend upon the data, in which case we call the estimate a **variable partition estimate.** One can easily verify that if all the $A_i$'s have finite Lebesgue measure, then $f_n$ integrates to one.

In 1956, Rosenblatt proposed the **moving window estimate**

$$f_n(x) = \frac{\mu_n(S(x,h))}{\lambda(S(x,h))},$$

where $S(x,h)$ is the ball of radius $h$ centered at $x$. This radius is also called the **smoothing factor.** On the real line, $f_n(x)$ counts the number of data points in $[x-h, x+h]$ and divides this number by $2hn$. Note that for large $n$ and fixed $h$, $f_n$ is close to

$$\frac{\mu(S(x,h))}{\lambda(S(x,h))},$$

which in turn is close to $f(x)$ if $h$ is small enough. Again, we will have to find a good value for $h$ as a function of $n$. Typical drawings of $f_n$ have discontinuities that are reminiscent of the histogram estimate. In fact, since $f_n$ has zero derivative almost everywhere, it can be considered as a variable partition histogram estimate. Users often demand smooth eye-pleasing estimates. The theory will support this common sense request, as we will see further on. The simplest way to generalize the moving window estimate is by replacing the window (which can be considered as a uniform density) by a general function $K$, called a **kernel:**

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i),$$

where

$$K_h(x) = \frac{1}{h^d} K\left(\frac{x}{h}\right).$$

This estimate is known as the **kernel estimate** or the Parzen-Rosenblatt estimate (Parzen (1962), Cacoullos (1966)).

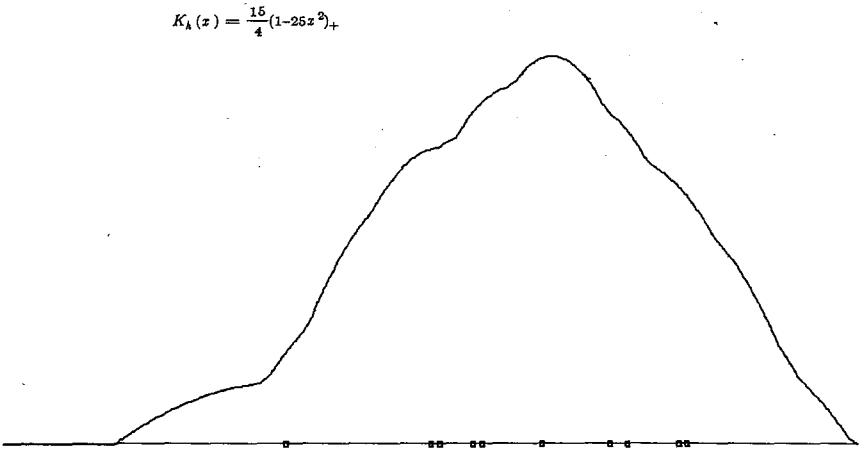


**Figure 2.2.**
Kernel estimate with $n = 10$. Ten data points are shown on axis.

Again, it is easy to verify that $f_n$ is a density when $K$ is a density. Remarkably, $f_n$ remains a density even if $K$ and/or $h$ depend upon the data (but not upon $x$). Nearly every desirable property of the histogram estimate is shared by the kernel estimate, so we won't have to bother too much with histogram estimates.

The derivative can be approximated in yet another way, by making the radius of the ball depend upon $x$ and the data:

$$f_n(x) = \frac{\mu_n(S(x, R_k))}{\lambda(S(x, R_k))},$$

where $R_k$ is the minimal value for which $\mu_n(S(x, R_k)) = k/n$. Thus, the ball centered at $x$ has radius equal to the distance between $x$ and its $k$-th nearest neighbor in the data sequence. First introduced by Fix and Hodges in 1951 (see also Loftsgaarden and Quesenberry, 1965), this estimate is called the **nearest neighbor estimate.**

We have

$$f_n(x) = \frac{k/n}{\lambda(S(x, R_k))}$$
$$\approx \frac{\mu(S(x, R_k))}{\lambda(S(x, R_k))} \quad (k \text{ large})$$

$$\approx \frac{d\mu}{d\lambda} = f(x) \quad (\frac{k}{n} \text{ small}) .$$

For good performance, we will once again need to balance two conflicting requirements: $k$ should be large, but $k/n$ should be small. The key property of the nearest neighbor estimate is that $\int f_n = \infty$, so that it is impossible to study its properties in $L_1$. To see this for the case $d = 1$, note that

$$f_n(x) \geq \frac{k}{n(x - X_{(k)})} , \quad x > X_{(1)} ,$$

where $X_{(1)} > \cdots$ are the order statistics. See figure 2.3 below.

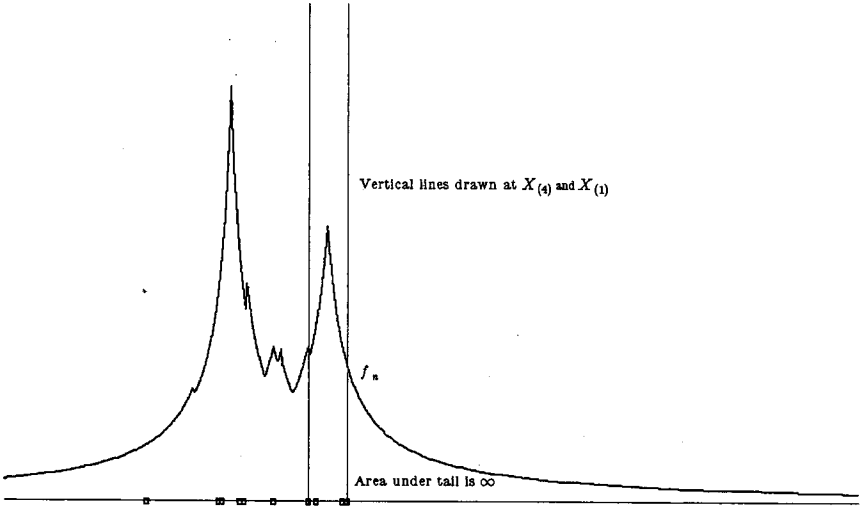


**Figure 2.3.**
Nearest neighbor estimate with $n = 10, k = 4$. Data points are shown on axis.

## 2.3. DIFFERENTIATION OF INTEGRALS.

In the study of the kernel estimate, we will need to know about the properties of the convolution (or smoothing) operator $*$. For two functions $f, g$ in $L_1$, we have

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

We begin with

**Theorem 2.1. Young's inequality.**
If $f, g \in L_1$, then $\int |f*g| \leq \int |f| \int |g|$.

**Proof of Theorem 2.1.**

$$\int \left| \int f(y)g(x-y)\,dy \right| dx \leq \iint |f(y)|\,|g(x-y)|\,dy\,dx$$
$$= \int |f(y)| \int |g(x-y)|\,dx\,dy = \int |f| \int |g| . \blacksquare$$

It is also necessary to recall from real analysis the following fact:

**Theorem 2.2.**
The continuous densities with compact support form a dense subclass in the class of all densities. In other words, for every $\epsilon > 0$, and for every density $f$, there exists a continuous density $g$, of compact support, such that $\int |f-g| \leq \epsilon$.

This theorem is just a special case of a more general theorem which states that the continuous compact support functions are dense in $L_1$. The consistency of density estimates can often be obtained by first proving the consistency for a dense subclass of nice densities, and then invoking Theorem 2.2.

**Theorem 2.3.**

Every kernel $K$ with $\int K = 1, \int |K| < \infty$ is an **approximate identity**, i.e. for every $f \in L_1$, we have

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

**Proof of Theorem 2.3.**

Assume first that the statement is true for a dense subspace of functions $g$. Then, for arbitrary $f \in L_1$,

$$\begin{aligned} \int |f * K_h - f| &\leq \int |f - g| * |K_h| + \int |f - g| + \int |g * K_h - g| \\ &\leq (\int |K| + 1) \int |f - g| + o(1) . \end{aligned}$$

The first term on the right-hand-side of this expression can be made as small as desired by choice of $g$, and the finiteness of $\int |K|$. Thus, we need only prove the $L_1$ convergence for a dense subclass, such as the class of continuous functions with compact support.

Define

$$L = KI_{||x|| \leq M}$$

for some large constant $M$. Let $A$ be $\{z : ||z - x|| \leq M$ for some $x \in$ Support $(f)\}$. Let us furthermore introduce the **modulus of continuity**

$$\omega(g, h) = \sup_{y, x : ||y|| \leq h} |g(x - y) - g(x)| .$$

Then the following chain of inequalities is valid:

$$\begin{aligned} \int |g * K_h - g| &= \int |g * K_h - g \int K_h| \\ &\leq \int |g * L_h - g \int L_h| + \int |g| \int |K_h - L_h| + \int |g * (K_h - L_h)| \\ &\leq \iint_A |g(x - y) - g(x)| \, |L_h(y)| \, dy \, dx + 2 \int |K - L| \\ &\leq \omega(g, Mh) \iint_A |L_h(y)| \, dy \, dx + 2 \int |K - L| \\ &\leq \omega(g, Mh) \lambda(A) \int |L| + 2 \int |K - L| \\ &= o(1) + 2 \int |K - L| \end{aligned}$$

and this can be made as small as desired by our choice of $M$. ■

## 2.4. CHARACTERISTIC FUNCTIONS.

If $f$ and $K$ are densities with characteristic functions $\phi$ and $\psi$ respectively, then the convolution $f*K_h$ has characteristic function $\psi(th)\phi(t)$. This observation allows us to deduce quite a bit of information. For example, by the uniqueness of characteristic functions, we see that

$$\int | f*K_h - f | = 0 \text{ if and only if } \psi(th)\phi(t) = \phi(t) \text{ for all } t .$$

But this is in turn equivalent to the condition that for all $t$, either $\psi(th)=0$ or $\phi(t)=0$. Since $\phi(t)\neq 0$ in a neighborhood of the origin, we need $\psi(th)=1$ for that same $t$-neighborhood. But $\psi$ can be one only at the origin. This forces $h$ to be zero. Thus, for any $h\neq 0$,

$$\int | f*K_h - f | > 0 .$$

Parseval's identity states that the integrals of the squares of functions are equal to the integrals of the squares of the Fourier transforms (or characteristic functions). Unfortunately, such a nice identity does not exist between the $L_1$ norm of a function on the real line and the $L_\infty$ norm of the corresponding Fourier transform. However, we have

**Theorem 2.4.**

Let $f$, $g$ be densities with characteristic functions $\phi$ and $\psi$. Then

$$\int | f - g | \geq \sup_t | \phi(t) - \psi(t) | .$$

**Proof of Theorem 2.4.**

$$| \phi(t) - \psi(t) | = | \int e^{itx} f(x)\, dx - \int e^{itx} g(x)\, dx |$$
$$\leq \int | e^{itx} | \, | f(x) - g(x) | \, dx = \int | f - g | . \blacksquare$$

Consequently, using $\psi$ again for the characteristic function of our kernel $K$, we have

$$\int | f*K_h - f | \geq \sup_t | \phi(t) | \, | 1 - \psi(th) | .$$

From this, we can deduce that if $h_n$ is a sequence of numbers for which

$$\lim_{n\to\infty} \int | f*K_{h_n} - f | = 0 ,$$

then $h_n \to 0$. For if this were not the case, then there would exist a subsequence $n'$ along which $h_{n'} \to H \in (0,\infty]$ as $n' \to \infty$. For $t \neq 0$, we would then have

$$\lim_{n' \to \infty} \psi(th_{n'}) = \begin{cases} \psi(tH) & (H < \infty) \quad \text{(continuity of } \psi\text{)} \\ 0 & (H = \infty) \quad \text{(Riemann-Lebesgue)} \end{cases}.$$

By the $L_1$–$L_\infty$ inequality, and the fact that $\phi(t) \neq 0$ in a neighborhood of the origin, we see that $\liminf_{n' \to \infty} \int | f * K_{h_{n'}} - f | > 0$, which is a contradiction.

## 2.5. INTEGRAL CONVERGENCE FROM POINTWISE CONVERGENCE.

The notion of pointwise convergence is stronger than that of integral convergence. This provides us with yet another tool for establishing the consistency of a density estimate. Scheffe (1947) first observed this interesting fact:

**Theorem 2.5. Scheffe's theorem.**

If $f_n$ is a sequence of densities (not estimates), and $f_n \to f$ almost everywhere where $f$ is a density, then $\int | f_n - f | \to 0$ as $n \to \infty$.

**Proof of Theorem 2.5.**

By the Lebesgue dominated convergence theorem,

$$\int | f_n - f | = 2 \int (f - f_n)_+ \to 0 . \blacksquare$$

Scheffe's theorem is not directly applicable to density estimates because the sequence $f_n$ needs to be deterministic. The void is filled by

**Theorem 2.6. Glick's theorem (Glick, 1974).**

If $f_n$ is a sequence of density estimates converging almost everywhere to a density in probability (or almost surely), then $\int | f_n - f | \to 0$ in probability (or almost surely).

**Proof of Theorem 2.5.**

The "in probability" half is proved by applying the Lebesgue dominated convergence theorem (LDCT) twice:

$$f \geq (f - f_n)_+ \to 0 \text{ in probability, a.e.}$$

implies, by LDCT,

$$f \geq E(f - f_n)_+ \to 0 \text{ a.e.}$$

and thus, again by LDCT,

$$E(\int | f - f_n |) = \int E(f - f_n)_+ \to 0 .$$

The almost sure part follows by a double application of Fubini's theorem:

$$P(\omega: f_n x) \not\to f(x)) = 0 \text{ for almost all } x\,(\lambda)$$
if and only if
$$\{(\omega, x): f_n(x) \not\to f(x)\} \text{ has } P \times \lambda \text{ measure } 0$$
if and only if
$$\lambda(x: f_n(x) \not\to f(x)) = 0 \text{ for almost all } \omega(P).$$

Thus, for almost all $\omega(P)$, we have $\int | f_n - f | \to 0$ by Theorem 2.4. ■

Pointwise convergence at almost all points usually requires a theorem in the spirit of the **Lebesgue density theorem**, stated here in a general form (see e.g. Wheeden and Zygmund, 1977):

**Theorem 2.7. The Lebesgue density theorem.**

Let $\mathbf{B}$ be a subclass of the Borel sets with the property that

$$\sup_{B \in \mathbf{B}} \frac{\lambda(B_0)}{\lambda(B)} \leq c < \infty$$

for some constant $c$, where $B_0$ is the smallest centered cube containing $B$ (see figure 2.4 below). Then, at almost all $x$,

$$\lim_{n \to \infty} \frac{\int_{x+B_n} |f(y) - f(x)| \, dy}{\lambda(B_n)} = 0$$

as $\lambda(B_n) \to 0$, where $B_n \in \mathbf{B}$ for all $n$. For the same $x$,

$$\lim_{n \to \infty} \frac{\int_{x+B_n} f(y) \, dy}{\lambda(B_n)} = \lim_{n \to \infty} \frac{\mu(x+B_n)}{\lambda(x+B_n)} = f(x).$$

These $x$ are called **Lebesgue points.**

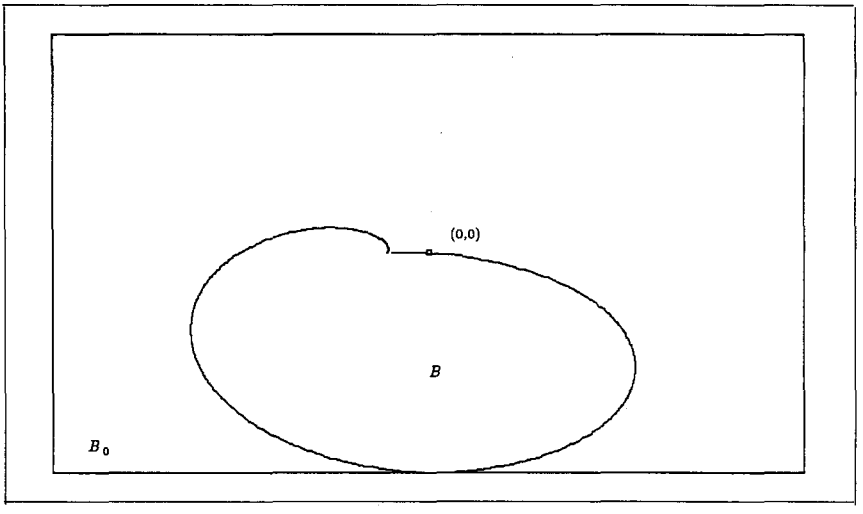


**Figure 2.4.**
$B_0$ is the smallest centered cube containing $B$

Theorem 2.7 is valid for shrinking balls and for shrinking centered hypercubes. Note that the last statement in the Theorem is equivalent to $f * g_n \to f$ a.e. where $g_n$ is the uniform density on $B_n$. We need something similar with $g_n$ replaced by $K_h$:

**Theorem 2.8.**

Let $K$ be an integrable function, satisfying

$$\int K = 1, \ |K| \le c < \infty, \ K = 0 \text{ off } S(0,M).$$

Then, for any density $f$,

$$f * K_h \to f \quad \text{at almost all } x$$

as $h \downarrow 0$.

**Proof of Theorem 2.8.**

$$\begin{aligned} |f * K_h - f| &= \left|\int (f(x-y) - f(x)) K_h(y)\, dy\right| \quad \left(\int K = 1\right) \\ &\le \int |f(x-y) - f(x)| \, |K_h(y)| \, dy \\ &\le \int_{S(0,Mh)} |f(x-y) - f(x)| \, ch^{-d} \, dy \\ &= ch^{-d} \, o(\lambda(S(0,Mh))) = o(1) \text{ as } h \downarrow 0. \ \blacksquare \end{aligned}$$

The conditions on $K$ in Theorem 2.7 are too strong. Stein (1970) has pointed out that one only needs $\int K = 1$ and an integrable radial majorant:

$$\int \sup_{||y|| \ge ||x||} |K(y)| \, dx < \infty.$$

## 2.6. EXERCISES.

2.1. Construct a sequence of densities $f_n$ and another density $f$ on [0,1] having the property that $\int |f_n - f| \to 0$ as $n \to \infty$, yet $\limsup_{n \to \infty} |f_n - f| > 0$ for all $x$.

2.2. Consider the nearest neighbor estimate in $R^d$ with $k = k_n$ varying in such a way that

$$\lim_{n\to\infty} k = \infty\,, \quad \lim_{n\to\infty} \frac{k}{n} = 0\,.$$

Show that $f_n \to f$ in probability for almost all $x$.

2.3. Construct a kernel $K$ with $\int |K| < \infty, \int K = 1$, for which the conclusion of Theorem 2.8 is false.

2.4. Consider the kernel estimate with kernel satisfying the conditions of Theorem 2.8. Assume that $h = h_n$ is such that $h \to 0$, $nh^d \to \infty$ as $n \to \infty$. Prove that $|f_n - f| \to 0$ in probability as $n \to \infty$ for almost all $x$. Conclude that $\int |f_n - f| \to 0$ in probability for all densities $f$. Hint: note that $E(f_n) = f * K_h$, and use an appropriate probability inequality for $|f_n - E(f_n)|$.

2.5. There is no converse to the inequality of Theorem 2.4. Construct densities $f, f_1, \ldots, f_n, \ldots$ with characteristic functions $\phi, \phi_1, \ldots, \phi_n, \ldots$ such that

$$\lim_{n\to\infty} \frac{\sup_t |\phi(t) - \phi_n(t)|}{\int |f - f_n|} = 0\,.$$

# Chapter Three
# CONSISTENCY OF THE KERNEL ESTIMATE

## 3.1. THE EQUIVALENCE THEOREM.

The object of this chapter is to illustrate various techniques for proving the consistency in $L_1$ of nonparametric estimates. We take as our main example the kernel estimate, because most of the problems encountered in practice can be illustrated clearly and simply.

The main result states that the kernel estimate is either consistent (in which case it converges in the strongest possible sense for all $f$ ) or not consistent (in which case it does not converge in any standard sense for one single $f$ ). There is no "inbetween".

**Theorem 3.1. (Devroye, 1983)**

Let $f_n$ be the kernel estimate with arbitrary density-kernel $K$, and let the smoothing factor $h$ depend upon $n$ only. Then the following statements are equivalent:

A. $\int |f_n - f| \to 0$ in probability for some $f$.

B. $\int |f_n - f| \to 0$ in probability for all $f$.

C. $\int |f_n - f| \to 0$ almost surely for all $f$.

D. For every $\epsilon > 0$, there exist $r, n_0 > 0$ (with $r$ independent of $f$, $K$) such that

$$P\left(\int |f_n - f| > \epsilon\right) \leq e^{-rn} \quad , n \geq n_0 \text{, all } f \ .$$

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

(E) implies (D) even if we allow negative-valued kernels, as long as $\int |K| < \infty$ , $\int K = 1$. But (A) does not generally imply (E) for these kernels.

We will take most of this chapter to prove this. Readers who are only interested in weak convergence are referred to a very short proof of weak convergence alluded to in exercise 2.4. Since (C) => (B) => (A), and since (D) => (C) by the Borel-Cantelli lemma, it suffices to prove (A) => (E), and (E) => (D). At the end of the chapter, we will extend Theorem 3.1 to the kernel estimate with data-dependent smoothing factor.

## 3.2. SIMPLE SPLITS INTO BIAS AND VARIATION.

The following trivial inequalities will be needed throughout:

$$\int |f_n - f| \leq \int |f * K_h - f| + \int |\mu_n * K_h - f * K_h|$$

(the first term is called the **bias**, and the second term is called the **variation**; we also used the fact that $f_n = \mu_n * K_h$),

$$E\left(\int |f_n - f|\right) \geq \int |f * K_h - f| .$$

(by Jensen's inequality, and the fact that $E(f_n) = f * K_h$), and

$$E\left(\int |f_n - f|\right) \geq \frac{1}{2} E\left(\int |\mu_n * K_h - f * K_h|\right) .$$

The last inequality follows from

$$E\left(\int |\mu_n * K_h - f * K_h|\right) \leq \int |f * K_h - f| + E\left(\int |f_n - f|\right) \leq 2E\left(\int |f_n - f|\right)$$

where we used the previous inequality. We first conclude that $E\left(\int |f_n - f|\right)$ tends to zero if and only if the bias term and the expected value of the variation term tend to zero.

We can also obtain the first half of (A) => (E): indeed, if (A) holds, then $E\left(\int |f_n - f|\right) \to 0$ for some $f$ (since $\int |f_n - f| \leq 2$). Hence, $\int |f * K_h - f| \to 0$. By a corollary of Theorem 2.4, $h \to 0$ as $n \to \infty$.

## 3.3. A LARGE DEVIATION INEQUALITY FOR THE MULTINOMIAL DISTRIBUTION.

When $K$ has compact support, it is easily seen that the behavior of $f_n(x)$ is nearly independent of the behavior of $f_n(y)$ if $x$ and $y$ are further than $ch$ apart for some constant $c$. Thus, the integral criterion $\int |f_n - f|$ sums very many nearly independent random variables, which is why we can expect to obtain some inequality like (D). One of the obstacles we have to deal with is the dependence due to the fact that the total sample size is $n$; this dependence is of

a multinomial nature. The tool needed in our proof is

**Theorem 3.2. A multinomial distribution inequality.**

Let $N_1, \ldots, N_k$ be a multinomial random vector with parameters $n, p_1, \ldots, p_k$. Then

$$P\left(\sum_{i=1}^{k} \left| \frac{N_i}{n} - p_i \right| \geq \epsilon\right) \leq 2^{k+1} e^{-n\epsilon^2/2}, \text{ all } \epsilon > 0 .$$

**Proof of Theorem 3.2.**

By Scheffe's theorem,

$$\sum_{i=1}^{k} \left| \frac{N_i}{n} - p_i \right| = 2 \sup_A \left| \frac{N(A)}{n} - P(A) \right|$$

where $A = \{\text{all } 2^k \text{ possible sets of integers from } 1,\ldots,k\}$, and $N(A)$ is the cardinality of $A$. By Bonferroni's inequality and Hoeffding's inequality (Hoeffding, 1963),

$$P\left( \sup_A \left| \frac{N(A)}{n} - P(A) \right| \geq \frac{\epsilon}{2} \right) \leq 2^k 2e^{-2n(\epsilon/2)^2} . \blacksquare$$

## 3.4. PROOF OF (E) => (D).

Theorem 2.4 states that $\int | f * K_h - f | \to 0$ when $h \to 0$ and $K$ is merely absolutely integrable with integral equal to one. Thus, it suffices to prove (D) for the variation only. This will be done in three steps, first for $K = \alpha I_R$ where $\alpha$ is a constant and $R$ is a rectangle, then for nonnegative $K$, and finally for absolutely integrable $K$.

For $K = \alpha I_R$, it helps to consider the partition $\Pi$ of the space into hypercubes of sides $h/N$, as shown below in figure 3.1.

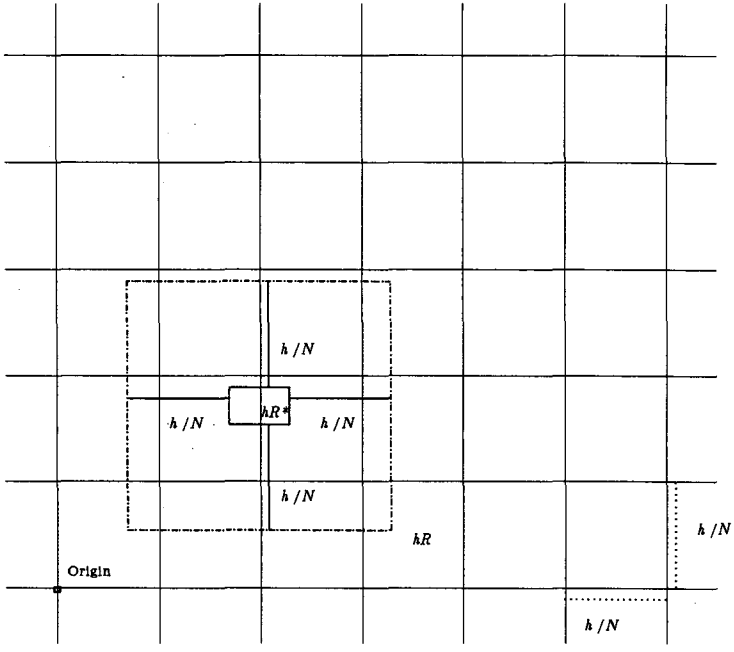


**Figure 3.1.**
Partition $\Pi$

We have

$$\int |\mu_n * K_h - f * K_h| = |\alpha| \int |\mu_n(x+hR) - \mu(x+hR)| h^{-d} \, dx .$$

Also, if we define $\text{Shell}_x$ as $x+hR$ minus the union of all sets $B \in \Pi$ for which $B \subseteq x+hR$, then $\text{Shell}_x \subseteq x+h(R-R^*)$. Thus,

$$|\mu_n(x+hR) - \mu(x+hR)| \leq \sum_{B \in \Pi,\, B \subseteq x+hR} |\mu_n(B) - \mu(B)| + (\mu+\mu_n)(\text{Shell}_x)$$

$$\leq \sum_{B \in \Pi,\, B \subseteq x+hR} |\mu_n(B) - \mu(B)| + (\mu+\mu_n)(x+h(R-R^*)) .$$

Collecting this yields

$$\int |\mu_n * K_h - f * K_h|$$

$$\leq |\alpha| h^{-d} \int \left( \sum_{B \in \Pi,\, B \subseteq x+hR} |\mu_n(B) - \mu(B)| + (\mu+\mu_n)(x+h(R-R^*)) \right) dx$$

$$\leq |\alpha| h^{-d} \left( \sum_{B \in \Pi} |\mu_n(B) - \mu(B)| \int_{B \subseteq x + hR} dx + 2\lambda(h(R - R^*)) \right)$$

(Young's inequality)

$$\leq |\alpha| \lambda(R) \sum_{B \in \Pi} |\mu_n(B) - \mu(B)| + 2|\alpha| \lambda(R - R^*). \quad \text{(Young's inequality)}$$

Let us take the last inequality as our starting point. For $\delta > 0$, and any density $K$, we can find a kernel $L$ of the form

$$L = \sum_{i=1}^{N_0} \alpha_i I_{R_i}$$

with the property that

$$\int |K - L| < \delta,$$

where $N_0$ is a constant, the $\alpha_i$'s are constants not exceeding some number $M$ in absolute value, and the $R_i$'s are disjoint finite rectangles. Note that $L = 0$ outside some hypercube $[-H, H]^d$. A standard triangular inequality yields

$$\int |\mu_n * K_h - f * K_h| \leq$$

$$\int |\mu_n * L_h - f * L_h| + \int \mu_n * |K_h - L_h| + \int f * |K_h - L_h|$$

$$\leq \sum_{i=1}^{N_0} |\alpha_i| h^{-d} \int |\mu_n(x + hR_i) - \mu(x + hR_i)| \, dx + 2\delta \quad \text{(Young's inequality)}$$

$$\leq \left( \sum_{i=1}^{N_0} |\alpha_i| \lambda(R_i) \right) \sum_{B \in \Pi} |\mu(B) - \mu_n(B)| + 2\delta + \sum_{i=1}^{N_0} 2|\alpha_i| \lambda(R_i - R_i^*).$$

The third term on the right-hand-side can be made smaller than $\delta$ by choosing $N$ large enough (each $R_i^*$ tends to $R_i$ as $N \to \infty$). The coefficient of the first term on the right-hand-side is equal to $\int |L| \leq 1 + \delta$. Thus, we have shown so far that for every $\delta > 0$, we can find $N$ large enough such that

$$\int |\mu_n * K_h - f * K_h| \leq 3\delta + (1 + \delta) \sum_{B \in \Pi} |\mu(B) - \mu_n(B)|$$

$$\leq 5\delta + \sum_{B \in \Pi} |\mu(B) - \mu_n(B)|.$$

$N$ depends upon $\delta$ and $K$ only, and $\Pi$ depends upon $h/N$. We are almost in a position now to utilize the multinomial inequality of Theorem 3.2, were it not for the fact that the partition $\Pi$ is infinite. Thus, it is necessary to "cut off" the tails of the distribution. To do this, consider the partition $\Pi$, and a finite partition $\Pi_r$, consisting of those sets of $\Pi$ that have a nonempty intersection with $[-r, r]^d$, where $r > 0$ is a constant to be picked further on. Let $\Pi_r^*$ be $\Pi_r \cup [-r, r]^{d\,c}$. See figure 3.2 below.

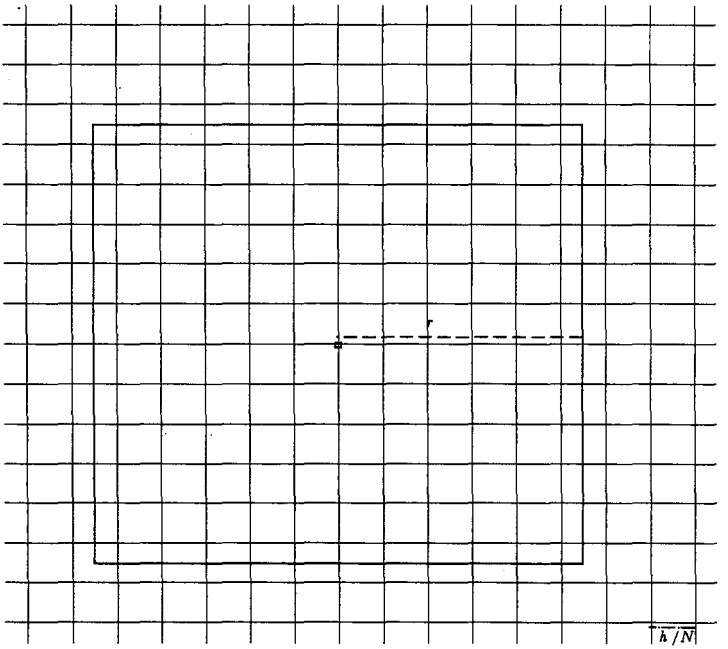


**Figure 3.2.**
Partitions $\Pi, \Pi_r$ .

The cardinality of $\Pi_r$ is at most

$$\left(\frac{2rN}{h}+2\right)^d = o(n) .$$

To take care of the tails, we argue as follows: let $T$ stand for the tail set, i.e. the complement of $[-r,r]^d$. Then

$$\sum_{B\in\Pi} |\mu(B)-\mu_n(B)| \le \sum_{B\in\Pi_r} |\mu(B)-\mu_n(B)| + \mu(T) + \mu_n(T)$$

$$\le \sum_{B\in\Pi_r} |\mu(B)-\mu_n(B)| + 2\,\mu(T) + |\mu(T)-\mu_n(T)|$$

$$\le \sum_{B\in\Pi_r^*} |\mu(B)-\mu_n(B)| + 2\,\mu(T) .$$

Now, $2\,\mu(T)$ can be made smaller than $\delta$ by choice of $r$. Recapitulating, this gives the inequality

$$\int | \mu_n *K_h - f *K_h | \leq 6\,\delta + \sum_{B \in \Pi^*_r} | \mu(B) - \mu_n(B) | .$$

where $r$ depends upon $\delta, f$ , and $N$ depends upon $\delta, K$. By the multinomial inequality of Theorem 3.2, for $\epsilon > 6\,\delta$, and $\rho \in (0,1)$,

$$P\left( \int | \mu_n *K_h - f *K_h | > \epsilon \right) \leq P\left( \sum_{B \in \Pi^*_r} | \mu(B) - \mu_n(B) | > \epsilon - 6\,\delta \right)$$

$$\leq 2^{2+(2+2rN/h)^d}\, e^{-\frac{1}{2} n (\epsilon - 6\delta)^2}$$

$$\leq e^{-(1-\rho)\frac{n\epsilon^2}{2}} , \quad n \geq n_0(\rho, \epsilon, K, f , \{h\}) .$$

This concludes the proof of (E) $\Rightarrow$ (D) for nonnegative $K$. Note that the inequality can be forced to hold for all $n, h$ with

$$n > \frac{16 + 4^{d+1}}{\rho \epsilon^2} ,$$

$$nh^d > \frac{4\, 2^d\, (2rN)^d}{\rho \epsilon^2} ,$$

if we pick

$$\delta = \frac{\epsilon}{6} \left( 1 - \sqrt{1 - \frac{\rho}{2}} \right) .$$

If $K$ can take negative values but is absolutely integrable with integral one, we have

$$\int | L | \leq \int | \, | L | - | K | \, | + \int | K |$$

$$\leq \int | L - K | + \int | K | < \delta + \int | K | .$$

This yields the inequality

$$\int | \mu_n *K_h - f *K_h | \leq 6\,\delta + \int | K | \sum_{B \in \Pi^*_r} | \mu(B) - \mu_n(B) | .$$

We can thus conclude that for every $\epsilon > 0$, $\rho \in (0,1)$, there exists an $n_0(\epsilon, \rho, K, f , \{h\})$, such that

$$P\left( \int | \mu_n *K_h - f *K_h | > \epsilon \right) \leq e^{-(1-\rho)\frac{n\epsilon^2}{2}} , \quad n \geq n_0 .$$

### 3.5. PROOF OF (A) => (E).

The necessity of the condition $h \to 0$ was already established above, so we need only be concerned with the condition $nh^d \to \infty$.

Observe first that

$$\int |f_n - f| \to 0 \text{ in probability}$$

$$\Rightarrow$$

$$E\left(\int |f_n - f|\right) \to 0$$

$$\Rightarrow$$

$$E\left(\int |\mu_n * K_h - f * K_h|\right) \to 0 .$$

Also,

$$E\left(\int |\mu_n * K_h - f * K_h|\right) \geq E\left(\int |\mu_n * L_h - f * L_h|\right) - 2\int |K - L| ,$$

where $L$ is a function close to $K$, defined by

$$L = \frac{\min(|K|, M)\operatorname{sign}(K)}{\int \min(|K|, M)\operatorname{sign}(K)} ,$$

i.e., $L$ is equal to $K$, truncated at $\pm M$. This inequality is obtained by applying Young's inequality twice:

$$\left| \int |\mu_n * K_h - f * K_h| - \int |\mu_n * L_h - f * L_h| \right|$$
$$\leq \int \mu_n * |K_h - L_h| + \int f * |K_h - L_h| = 2 \int |K - L| .$$

We will assume that $K$ has compact support, vanishing off $T = [-C/2, C/2]^d$. This is used in

$$E\left(\int |\mu_n * L_h - f * L_h|\right) \geq \int |f * L_h| \; P(x + hT \text{ is empty}) \, dx$$
$$= \int |f * L_h| \; (1 - \mu(x + hT))^n \, dx .$$

By Theorem 2.8, $f * L_h \to f$ almost everywhere as $h \to 0$ (since $L$ has compact support, is bounded, and integrates to one). Thus, as $h \to 0$, $\mu(x + hT) \sim (hC)^d f(x)$ almost everywhere. By Fatou's lemma and the standard inequality $1 - x \geq e^{-\frac{x}{1-x}}$, valid for $0 \leq x < 1$, we have, if $h \to 0$ as $n \to \infty$,

$$\liminf E\left(\int |\mu_n * L_h - f * L_h|\right) \geq \int f e^{-\limsup \frac{n\mu(x+hT)}{1-\mu(x+hT)}} .$$

Assume that along a subsequence, $nh^d \to s \in [0, \infty)$. Then $h \to 0$. The previous inequality can be applied to this subsequence in which case the right-hand-side should be replaced by

$$\int f e^{-sC^d f} .$$

This shows that

$$2\int |K-L| \geq \int f e^{-sC^d f} .$$

But because $M$ was arbitrary and $\inf_M 2\int |K-L| = 0$, we conclude that

$$\int f e^{-sC^d f} = 0.$$

This can only be the case if $s=\infty$, which is a contradiction. The proof is complete if we can extend it to the case of kernels with unbounded support. See exercise 3.2.

## 3.6. DATA-BASED SMOOTHING.

Assume that the smoothing factor $h$ is a Borel measurable function of the data. We will write $H$ instead of $h$ to stress the fact that the smoothing factor is a random variable. The main result of this section is

**Theorem 3.3.**

Let $K$ be an arbitrary density, and assume that

$H \to 0$ in probability (almost surely, completely),

$nH^d \to \infty$ in probability (almost surely, completely).

Then $\int |f_n - f| \to 0$ in probability (almost surely, completely).

The proof is based upon the fact that small changes in $h$ do not affect the estimate very much. This is captured in two simple lemmas:

**Lemma 3.1.**

For any density $K$,

$$\lim_{h \to 1} \int |K_h - K_1| = 0 .$$

**Proof of Lemma 3.1.**

When $K$ is continuous, $K_h \to K_1$ for all $x$. Hence, $\int |K_h - K_1| \to 0$ by Scheffe's theorem. For arbitrary $K$,

$$\begin{aligned}\int |K_h - K_1| &\leq \int |K_h - L_h| + \int |L_h - L_1| + \int |L_1 - K_1| \\ &= 2\int |K - L| + \int |L_h - L_1| ,\end{aligned}$$

where $L$ is a continuous density. The last term is $o(1)$ because $L$ is continuous. The first term can be made arbitrarily small because the continuous densities are dense in the space of all densities. ■

**Lemma 3.2.**

Let $f_{nh}$ be the kernel estimate with kernel $K$ ($K$ is a density), and with smoothing factor $h$. Then

$$\int |f_{nh} - f_{nh'}| \leq \phi\left(\frac{h'-h}{h}\right), \quad h' > h ,$$

where

$$\phi(\delta) \stackrel{\Delta}{=} \sup_{1-\delta \leq u \leq 1+\delta} \int |K_u - K| .$$

**Proof of Lemma 3.2.**

$$\int |\mu_n * K_h - \mu_n * K_{h'}| \leq \int |K_h - K_{h'}| = \int |K_{h'/h} - K_1| . \blacksquare$$

Lemmas 3.1 and 3.2 together show that small fluctuations in $h$ have small effects on $f_{nh}$. With this observation, we can establish a uniform inequality in the spirit of (D) of Theorem 3.1:

**Lemma 3.3. The key inequality.**

Let $H_n = [h'_n, h''_n]$ be a sequence of deterministic intervals, where $h''_n \to 0$ and $nh'^d_n \to \infty$ as $n \to \infty$. For every $\epsilon > 0$, there exist $n_0 > 0$ and $r > 0$ such that

$$P\left( \sup_{h \in H_n} \int | f_{nh} - f | > \epsilon \right) \leq e^{-rn\epsilon^2}, \quad n \geq n_0 .$$

The inequality states that after seeing the data, and knowing $f$, and choosing the worst $h$ in a certain set $H_n$ in function of this knowledge, the kernel estimate remains almost surely consistent in the $L_1$ sense.

**Proof of Lemma 3.3.**

The proof is based upon a reduction of the supremum over an uncountable set to a supremum over a finite set. The interval $H_n$ is partitioned as shown in figure 3.3.

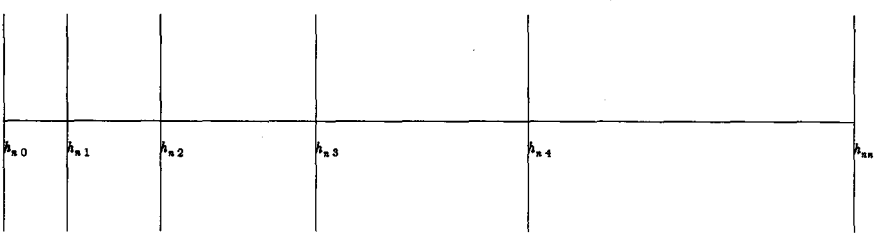


**Figure 3.3.**
Partition of $H_n = [h'_n, h''_n] = [h_{n0}, h_{nn}]$ into $n$ intervals.

The interval sizes in the partition grow geometrically, and the boundaries are defined by

$$h_{ni} = (1+\delta_n)^i h'_n \quad , 0 \leq i \leq n ,$$

where

$$(1+\delta_n)^n = \frac{h''_n}{h'_n} = \left( \frac{h''^d_n}{nh'^d_n} \right)^{\frac{1}{d}} n^{\frac{1}{d}} .$$

Observe that $\delta_n \to 0$. For $1 \le i \le n$, we have

$$\sup_{h_{n\,i-1} \le h \le h_{ni}} \int | f_{nh} - f_{nh_{n\,i-1}} | \le \sup_{h_{n\,i-1} \le h \le h_{ni}} \phi\left(\frac{h}{h_{n\,i-1}} - 1\right) \text{ (Lemma 3.2)}$$

$$\le \phi\left(\frac{h_{ni}}{h_{n\,i-1}} - 1\right) = \phi(\delta_n) \text{ (definition of } \phi \text{ and } \delta_n)$$

$$< \epsilon \text{ for all } n \text{ large enough, since } \delta_n \to 0 .$$

Thus, we have

$$\sup_{h \in H_n} \int | f_{nh} - f | \le \sup_{1 \le i \le n} \left( \int | f_{nh_{n\,i-1}} - f | + \sup_{h_{n\,i-1} \le h \le h_{ni}} \int | f_{nh} - f_{nh_{n\,i-1}} | \right)$$

$$\le \sup_{1 \le i \le n} \int | f_{nh_{n\,i-1}} - f | + \epsilon \text{ (all } n \ge n_1) .$$

By Bonferroni's inequality,

$$P\left( \sup_{h \in H_n} \int | f_{nh} - f | > 2\epsilon \right) \le \sum_{i=1}^{n} P\left( \int | f_{nh_{n\,i-1}} - f | > \epsilon \right)$$

$$\le n \, e^{-\frac{1-\rho}{2} n \epsilon^2}$$

when $n \ge n_1$, $n \ge c_1$, $nh'^{\,d}_n \ge c_2$ and $h''_n \le c_3$ for some positive constants $c_j$ obtained in the proof of Theorem 3.1 in the previous section. The constant $\rho \in (0,1)$ is picked by the user and affects these constants. Thus, there exists a constant $n_0$ such that the right-hand-side of the last chain of inequalities does not exceed

$$e^{-\frac{1-2\rho}{2} n \epsilon^2}, \quad n \ge n_0 .$$

This concludes the proof of Lemma 3.3. ■

**Proof of Theorem 3.3.**

First observe that

$$I_{[h + (nh^d)^{-1} \ge \epsilon]} \to 0 \text{ completely}$$

implies

There exists an $\epsilon'_n \downarrow 0$
for which $I_{[h + (nh^d)^{-1} \ge \epsilon'_n]} \to 0$ completely.

Choose $H_n = [h'_n, h''_n]$ where

$$h''_n = \epsilon_n , \quad h'_n = (n \epsilon_n)^{-1/d} , \quad \epsilon_n = \max(\epsilon'_n , n^{-1/(d+1)}) .$$

Verify the following statements:

A. $h''_n \to 0$.

B. $nh'_n{}^d = 1/\epsilon_n \to \infty$.

C. $I_{[h+(nh^d)^{-1} \geq \epsilon_n]} \to 0$ completely.

D. $h'_n \leq h''_n$ (because $(n\epsilon_n)^{-1} \leq \epsilon_n{}^d$).

E. $I_{h \notin H_n} \to 0$ completely.

Statement (E) follows from (C) and

$$I_{[h+(nh^d)^{-1} \geq \epsilon_n]} = \frac{1}{2}\left(I_{[h+(nh^d)^{-1} \geq h''_n]} + I_{[h+(nh^d)^{-1} \geq (nh'_n{}^d)^{-1}]}\right)$$

$$\geq \frac{1}{2}\left(I_{[h \geq h''_n]} + I_{[h \leq h'_n]}\right) = \frac{1}{2} I_{[h \notin H_n]} .$$

From (E) and Lemma 3.3, we have

$$I_{[\int |f_n - f| > \epsilon]} \leq I_{[h \leq h'_n]} + I_{[\sup_{h \in B_n} \int |f_{nh} - f| > \epsilon]} \to 0 \text{ completely} .$$

Note that we can replace the word "completely" in the proof by "almost surely" or "in probability". ■

## 3.7. EXERCISES.

3.1. Consider a histogram estimate on the real line defined by the partition $[ih, (i+1)h)$, for $i = \ldots, -2, -1, 0, 1, 2, \ldots$. Here $h$ plays the role of a smoothing factor. In the notation of Theorem 3.1, prove that (E) $\Rightarrow$ (D).

3.2. Extend the proof of (A) $\Rightarrow$ (E) in Theorem 3.1 to cover the case of kernels with unbounded support.

# Chapter Four
# ROBUSTNESS

**4.1. DEFINITION.**

An estimator is robust if small changes in the underlying distribution induce small changes in the estimator. In other words, the estimator is not hypersensitive to the distribution. One possible definition of **robustness**, adapted from Bickel (1976), states that a density estimate $f_n$ is robust at $f$ if

$$\sup_{n \geq 1,\, \epsilon > 0} \left\{ \frac{\sup_{g \in S(f,\epsilon)} E_g \int | g_n - g |}{\epsilon + E_f \int | f_n - f |} \right\} \leq C < \infty$$

for some constant $C$. Here $S(f,\epsilon)$ is the $L_1$ ball of radius $\epsilon$ centered at $f$. The notation $E_f$ denotes the expected value with respect to a sample of size $n$ drawn from $f$. The function $g_n$ is identical to $f_n$, but the distinction is made to stress the fact $g_n$ uses data drawn from $g$.

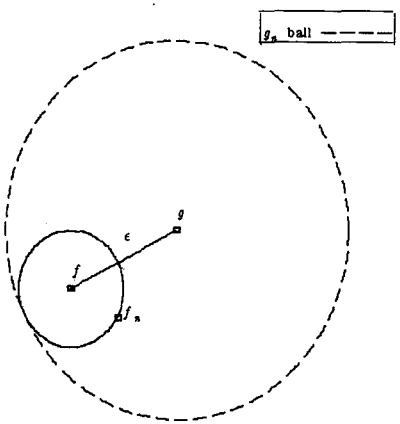


**Figure 4.1a.**
Crude illustration of size of region in which $g_n$ is allowed to fall.

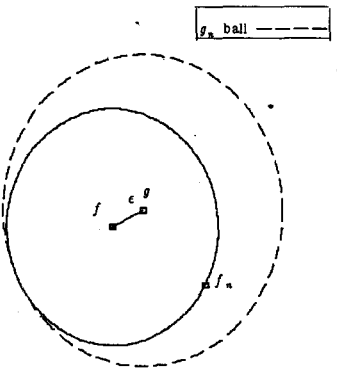


**Figure 4.1b.**
Crude illustration of size of region in which $g_n$ is allowed to fall.

In figure 4.1, we roughly illustrate what this means. Being a bit sloppy in our statement, we might say that $g_n$ should be in a ball centered at $g$ of radius not exceeding $C(\int | f-g | + \int | f_n-f | )$. When $g$ is very close to $f$, the second term dominates, while for $g$ far away from $f$, the $\int | f-g |$ term dominates. $C$ could be called a coefficient of elasticity, as it reflects how a sudden move of size $\int | f-g |$ can influence a move of the estimate, which is of size $\int | f_n-g_n |$.

## 4.2. AN EXAMPLE: A PARAMETRIC ESTIMATE.

To illustrate the definition of robustness given in the previous section (which is by no means the only possible definition), consider the exponential density $f(x) = e^{-x}$, $x > 0$. The estimator $f_n$ is the standard parametric estimator, i.e. it is an exponential density with parameter equal to the sample mean

$$\frac{1}{n} \sum_{i=1}^{n} X_i .$$

Consider a density $g \in S(f, \epsilon)$ constructed as follows:

$$g = (1 - \frac{\epsilon}{2}) f + \frac{\epsilon}{2} \psi_M$$

where $\psi_M$ is a spike uniform density function of height $M$ and width $1/M$ with support on $[M, M + 1/M]$. Note that indeed, $\int | g - f | \leq \epsilon$.

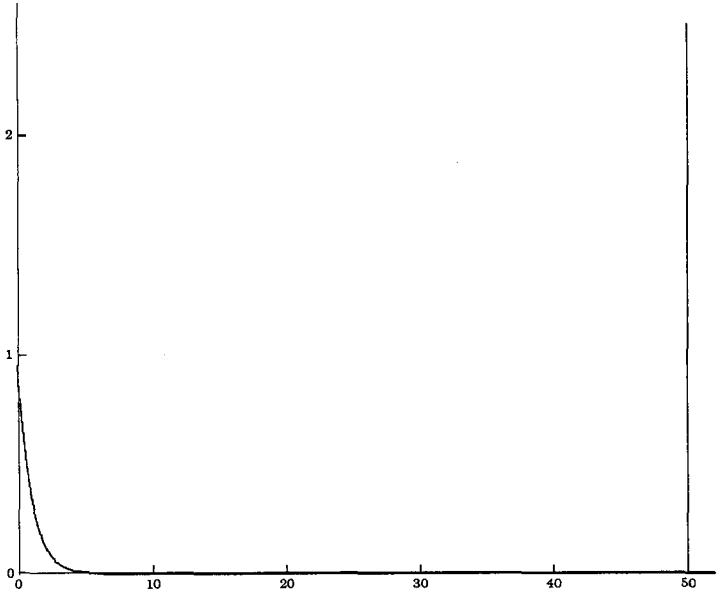


**Figure 4.2.**

Mixture $g = 0.95\, e^{-x} + 0.05\, \psi_{50}$ where $\psi_{50}$ is the uniform density on [50,50.02]

Clearly, $g_n$ is exponential with parameter at least equal to $M/n$ if one of the $X_i$'s is drawn from the $\psi_M$ part of the mixture. Thus,

$$\sup_{g \in S(f,\epsilon)} E_g \int | g_n - g | \geq 2 P(\text{Binomial}(n, \frac{\epsilon}{2}) > 0)$$

$$= 2 \left( 1 - (1 - \frac{\epsilon}{2})^n \right) \to 2 \text{ as } n \to \infty .$$

By the weak law of large numbers, we know that the sample mean converges in probability to one. Therefore, by continuity of the exponential density, $f_n$ tends to $f$ in probability at all $x$, and thus $E_f \int | f_n - f | \to 0$ by Glick's theorem. This implies that

$$\sup_{n \geq 1} \left\{ \frac{\sup_{g \in S(f,\epsilon)} E_g \int | g_n - g |}{\epsilon + E_f \int | f_n - f |} \right\} \geq \frac{2}{\epsilon} .$$

Therefore, the estimator is not robust.

## 4.3. THE KERNEL ESTIMATE.

The kernel estimate is robust at all $f$, and for all choices of $h$ and $K$. In fact, we will prove the following inequality:

**Theorem 4.1.**

Let $f_n$ be the kernel estimate with absolutely integrable kernel $K$, and arbitrary smoothing parameter $h > 0$. Then, for all $n \geq 1, \epsilon > 0$ and $f$,

$$\frac{\sup_{g \in S(f,\epsilon)} E_g \int | g_n - g |}{\epsilon + E_f \int | f_n - f |} \leq 1 + \int | K | .$$

**Proof of Theorem 4.1.**

The numerator of the left-hand-side in the inequality is bounded from above by in the usual manner:

$$E_g \int | g_n - g |$$

$$\leq E_f \int | f_n - f | + \int | f - g | + E_{f,g} \int | f_n - g_n | .$$

$$\leq E_f \int |f_n - f| + \epsilon + E_{f,g} \int |f_n - g_n| .$$

Note that the expected value $E_{f,g}$ is with respect to two samples. Nothing keeps us however from making these samples depend upon each other. In fact, we will show that there exists a probability space (an embedding) such that

$$E_{f,g} \int |f_n - g_n| \leq \int |K| \int |f - g| \leq \epsilon \int |K| \quad (\text{since } g \in S(f,\epsilon)) .$$

Once we have shown that, we have

$$\frac{\sup_{g \in S(f,\epsilon)} E_g \int |g_n - g|}{\epsilon + E_f \int |f_n - f|} \leq 1 + \frac{1}{\epsilon} \sup_{g \in S(f,\epsilon)} E_{f,g} \int |g_n - f_n|$$

$$\leq 1 + \int |K| .$$

This would then conclude the proof of the Theorem. ■

**The embedding device.**

The object here is to construct two dependent samples of size $n$ each, one drawn from $f$, and one drawn from $g$, such that

$$E_{f,g} \int |f_n - g_n| \leq \int |f - g| \int |K| .$$

Observe that there is no hope of obtaining this with two independent samples, for as $h \to 0$, the $L_1$ distance between $f_n$ and $g_n$ tends to 2 almost surely. The construction of the samples can be done as follows (see Devroye, 1985): define

$$\Delta = \frac{1}{2} \int |f - g| .$$

Then define the following densities:

$$f_{\min} = \frac{\min(f,g)}{1-\Delta} ,$$

$$f_0 = \frac{f - \min(f,g)}{\Delta} ,$$

$$g_0 = \frac{g - \min(f,g)}{\Delta} .$$

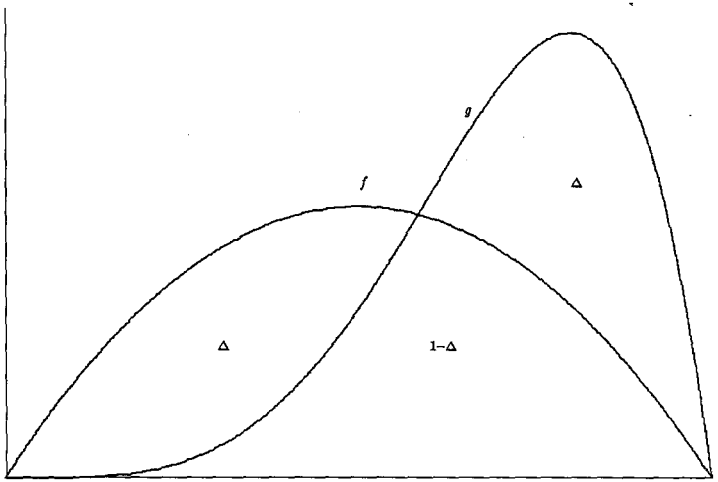


**Figure 4.3.**

$f$ and $g$ are shown. Samples are drawn from densities proportional to $(f-g)_+$, $(g-f)_+$ and $\min(f, g)$.

Four independent samples of iid random vectors are considered:

$$U_1, \ldots, U_n \sim f_{\min}$$

$$V_1, \ldots, V_n \sim f_0$$

$$W_1, \ldots, W_n \sim g_0$$

$$Z_1, \ldots, Z_n \sim \text{Bernoulli}(\Delta).$$

Then, define

$$X_i = X'_i = U_i \quad \text{if } Z_i = 0,$$
$$X_i = V_i, X'_i = W_i \quad \text{if } Z_i = 1, 1 \le i \le n.$$

We claim that

$$(X_1, \ldots, X_n)$$

is an iid sample drawn from $f$, and that

$$(X'_1, \ldots, X'_n)$$

is an iid sample drawn from $g$. This is based upon the mixture decomposition

$$f = (1-\Delta)f_{\min} + \Delta f_0 .$$

What matters is that the $X_i$'s and the $X'_i$'s agree except in $N$ components, where $N$ is binomial $(n,\Delta)$. Let $E$ be the expected value with respect to the probability measure defined above. Then

$$E\int |f_n - g_n| = E\int \left| \frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i) - \frac{1}{n}\sum_{i=1}^{n} K_h(x-X'_i) \right| dx$$

$$\leq E\int \left| \frac{1}{n}\sum_{i=1}^{N} K_h(x-V_i) \right| dx + E\int \left| \frac{1}{n}\sum_{i=1}^{N} K_h(x-W_i) \right| dx$$

$$\leq E\left(\frac{2N}{n}\right) \int |K| = \int |f-g| \int |K| . \blacksquare$$

## 4.4. APPLICATION: BERAN'S ROBUST PARAMETRIC ESTIMATES.

Robust nonparametric estimates such as the kernel estimate can be used to define robust parametric estimates. In this section, we follow a suggestion of Beran (1977). Let $\theta$ be a parameter and let $\mathbf{F} = \{f_\theta ; \theta \in \Theta\}$ be a family of densities parametrized by $\theta$. For example, $\theta$ can be thought of as the vector $(\mu,\sigma^2)$ defining the location and scale of a normal family. If $f_n$ is a suitable nonparametric estimate of $f$, then it is possible to estimate $\theta$ by $\theta_n$ where $\theta_n$ is the solution of the following optimization problem:

$$H_2(f_{\theta_n}, f_n) = \inf_{\theta \in \Theta} H_2(f_\theta, f_n) .$$

In our $L_1$ setting, we could define $\theta_n$ by the optimization problem

$$L_1(f_{\theta_n}, f_n) = \inf_{\theta \in \Theta} L_1(f_\theta, f_n) .$$

In other words, $f_{\theta_n}$ is closest to $f_n$ in $\mathbf{F}$.

**Theorem 4.2.**

The $L_1$ version of Beran's estimate is robust at all $\theta \in \Theta$, provided that it is based upon an estimator $f_n$ that is robust at all $f$.

**Proof of Theorem 4.2.**

Let $f = f_\theta$ for some $\theta \in \Theta$. Then

$$\sup_{g \in S(f,\epsilon)} E_g \int |f_{\theta_n} - g| \leq \sup_{g \in S(f,\epsilon)} \left( E_g \int |f_{\theta_n} - g_n| + E_g \int |g - g_n| \right) .$$

The last term in the inequality does not exceed $C(\epsilon + E_f |f_n - f|)$ if $f_n$ is robust at $f$, and has robustness constant $C$ (see definition of robustness). The term in front of it is bounded easily:

$$E_g \int |f_{\theta_n} - g_n| \leq E_g \int |g_n - g| + \inf_\theta \int |f_\theta - g| \leq E_g \int |g_n - g| + \epsilon .$$

Therefore,

$$\sup_{g \in S(f,\epsilon)} E_g \int |f_{\theta_n} - g| \leq 2C\left(\epsilon + E_f |f_n - f|\right) + \epsilon$$
$$\leq (2C+1)\left(\epsilon + E_f |f_n - f|\right) . \blacksquare$$

## 4.5. EXERCISES.

4.1. Prove or disprove: the histogram estimate based upon the partition $[ih, (i+1)h)$, for $i = 0, \pm 1, \pm 2, \ldots$ and fixed $h$ is robust at all $f$.

4.2. Let $M_n$ be the median of $(X_1, \ldots, X_n)$, a sample of iid random variables drawn from some density. Let $f$ be the normal (0,1) density, and let $f_n$ be the normal $(M_n, 1)$ density. Is $f_n$ robust at $f$? Prove or disprove.

4.3. Consider a kernel estimate with kernel $K$ and data-based smoothing factor $h$. Verify the robustness, at all $f$, of this estimate when

(i) $h$ is a function of $n$ times the difference between the $n/3$ and $2n/3$ quantiles in the data.

(ii) $h$ is a constant (depending upon $n$ only) times the standard sample based estimate of the standard deviation (i.e. $h^2 = (c/n)\sum (X_i - \hat{\mu})^2$ where $\hat{\mu}$ is the sample average).

4.4. Show that Beran's estimate $f_{\theta_n}$ of section 4.4 is consistent for all $f \in \mathbf{F}$ when in the definition of his estimate, one uses a consistent estimate $f_n$.

# Chapter Five
# MINIMAX BOUNDS

## 5.1. MINIMAX THEORY.

Minimax theory is concerned with the quantity

$$m(n,\mathbf{F}) = \inf_{f_n} \sup_{f \in \mathbf{F}} E\left(\int |f_n - f|\right),$$

where $\mathbf{F}$ is a suitable class of densities and $f_n$ is any estimator based upon an iid sample of size $n$ drawn from $f$.

**Lower bounds** for $m(n,\mathbf{F})$ are very important, because they tell us about the minimal expected error committed by any density estimate on at least one member of $\mathbf{F}$. They could for example be used to determine whether a certain sample size $n$ suffices to obtain a given expected error with some estimate, for the class of densities under investigation. In this chapter, we will exclusively deal with lower bounds. Wherever possible, we will also be concerned with the values of the constants in the lower bounds.

Ideally, one would like to know $m(n,\mathbf{F})$ exactly, but this is often difficult to compute. There are methods for determining upper bounds without actually constructing estimators. In chapter 6, a straightforward constructive method is developed. For large classes $\mathbf{F}$, good upper bounds can also be obtained by analyzing the performance of one of the popular nonparametric estimates (see e.g. chapter 7 on the kernel estimate). The real issue of course is the construction of a **minimax-optimal estimate** $f_n$, i.e. an estimate for which

$$\sup_{f \in \mathbf{F}} E\left(\int |f_n - f|\right) \leq C\ m(n,\mathbf{F})$$

for some universal constant $C$.

Here is a partial list of some possible classes $\mathbf{F}$:

| | |
|---|---|
| $L_1$ | all densities |
| $M_B$ | all monotone densities on [0,1] bounded by $B$ |
| $B_c$ | all densities on [0,1] bounded by $c$ |
| $U$ | all unimodal densities with a mode at 0 |
| $LC$ | all log-concave densities with a mode at 0 |
| $W(s,\alpha,C)$ | all densities on [0,1] with $s-1$ absolutely continuous derivatives and $f^{(s)}(x)-f^{(s)}(y) \leq C \mid x-y \mid^{\alpha}$ |
| $NS$ | all normal scale mixtures |
| $N$ | all normal densities with mode at 0 |

A word of warning. Designing estimates for good minimax performance over a set $\mathbf{F}$ could have negative side-effects. As we will see, the estimates may not always be consistent for $f \notin \mathbf{F}$, or the rate of convergence for certain $f$ in $\mathbf{F}$ could be inferior compared to that of other estimates. One common complaint is that $m(n,\mathbf{F})$ has little or no information about $m(n,\mathbf{G})$ for small classes of densities $\mathbf{G}$ properly contained in $\mathbf{F}$.

We will illustrate three techniques for obtaining lower bounds:

A. The "low-probability method". This method is based upon Devroye (1983).

B. Information-theoretic methods based upon the work of Assouad (1983), Bretagnolle, Huber (1979) and Birge(1986). These make use of the $H_2$ distance or the Kullback-Leibler numbers.

C. Methods based upon reductions to sufficient statistics.

## 5.2. THE LOW-PROBABILITY METHOD.

The first technique is applicable when $\mathbf{F}$ is a rich family of densities. One first partitions the space $R^d$ into very many sets $A_i$, and assigns a probability $p_i$ to each set $A_i$ in the partition, where $\sum_i p_i = 1$. On $A_i$, we define two nonnegative functions $g_i, h_i$ such that

$$\int g_i = \int h_i = p_i ,$$
$$\int | g_i - h_i | = 2p_i .$$

These conditions force $g_i, h_i$ to be disjoint. Next, define a subclass parametrized by

$$\theta = 0.\theta_1\theta_2\theta_3\theta_4 \cdots \quad \text{(binary expansion)}$$

with as a typical member

$$f_\theta(x) = \begin{cases} g_i(x), & x \in A_i, \ \theta_i = 1 \\ h_i(x), & x \in A_i, \ \theta_i = 0 \end{cases}.$$

In other words, the bit $\theta_i$ is used to choose between $g_i$ and $h_i$ on $A_i$. We require of course that $\{f_\theta\}$ be a subclass of $\mathbf{F}$. Usually, this can only be done when $\mathbf{F}$ is exceptionally vast. Three examples follow.

**Theorem 5.1.**

If $\mathbf{F}$ is a class containing a subclass of the type described above (for any probability vector $(p_1, p_2, \ldots)$), then

$$\inf_{f_n} \sup_{f \in \mathbf{F}} E\left(\int | f_n - f |\right) \geq 1 .$$

**Proof of Theorem 5.1.**

The proof uses the following construction: consider $n$ iid random variables drawn from each of the densities $g_i/p_i$, $h_i/p_i$, and let all samples be independent. Consider furthermore three other independent random variables or vectors, $\Theta$, $(\sigma_1, \ldots, \sigma_n)$, and $(N_1, N_2, \ldots)$, where $\Theta$ is uniformly distributed on $[0,1]$, $(\sigma_1, \ldots, \sigma_n)$ is a uniform random permutation of $(1, \ldots, n)$, and $(N_1, N_2, \ldots)$ is multinomially distributed with parameters $(n; p_1, p_2, \ldots)$. Define $(X_1, \ldots, X_n)$ as $(Y_{\sigma_1}, \ldots, Y_{\sigma_n})$ (a random permutation of the $Y_i$'s) where the $Y_i$'s contain the first $k$ random variables of the $g_i/p_i$ sequence if $N_i = k$ and $\Theta_i = 1$, and the first $k$ random variables of the $h_i/p_i$ sequence if $N_i = k$ and $\Theta_i = 0$. Observe that the $X_i$'s form an iid sample drawn from $f_\Theta$ (given $\Theta$). Furthermore, on $N_i = 0$, $\Theta_i$ and $X_1, \ldots, X_n$ are conditionally independent. We argue as follows:

$$\sup_{f\in\mathbf{F}} E(\int |f_n - f|)$$

$$\geq \sup_{\theta} E(\int |f_n - f_\theta|) \quad \text{(subclass of F)}$$

$$\geq E(\int |f_n - f_\Theta|) \quad \text{(randomization)}$$

$$= E\left(\sum_i \int_{A_i} \left\{ |f_n - g_i| I_{\Theta_i=1} + |f_n - h_i| I_{\Theta_i=0} \right\}\right)$$

$$\geq E\left(\sum_i I_{[N_i=0]} \int_{A_i} \left\{ |f_n - g_i| I_{\Theta_i=1} + |f_n - h_i| I_{\Theta_i=0} \right\}\right)$$

$$= E\left(\sum_i I_{[N_i=0]} \frac{1}{2} \int_{A_i} \left\{ |f_n - g_i| + |f_n - h_i| \right\}\right)$$

(conditional independence)

$$\geq E\left(\sum_i I_{[N_i=0]} \frac{1}{2} \int_{A_i} |g_i - h_i|\right)$$

$$= \frac{1}{2} \sum_i (2p_i)(1-p_i)^n$$

$$= \sum_i p_i (1-p_i)^n \ .$$

The proof is complete if we can show that

$$\sup_{p_1,p_2,\ldots:p_i\geq 0, \sum_i p_i = p} \sum_i p_i (1-p_i)^n = p \ .$$

This is most easily seen by taking $p_i = p/M$ for $1 \leq i \leq M$. Then

$$\sum_{i=1}^{M} p_i (1-p_i)^n = \left(1 - \frac{p}{M}\right)^n \to 1 \text{ as } M \to \infty \ . \blacksquare$$

## 5.3. EXAMPLES OF RICH CLASSES.

**$\mathbf{F} = B_2$ : the class of all densities on [0,1] bounded by 2.**

It should be clear that the subclass condition of Theorem 5.1 is applicable to $B_2$.

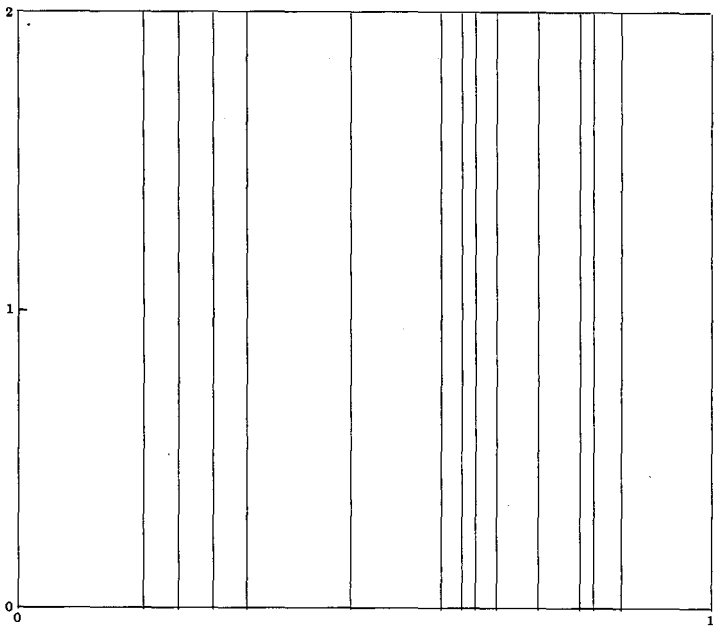


**Figure 5.1.**
A partition for $B_2$.

It suffices to take all the $g_i$'s and $h_i$'s equal to $2I_A$ for some set $A$; for $g_i$, $A$ is the leftmost half of $A_i$, and for $h_i$, it is the rightmost half of $A_i$.

**F consists of all densities with infinitely many absolutely continuous derivatives, and $|f^{(s)}| < C_s$ for $s = 0,1,2,\ldots$.**

Not all arbitrary sequences of constants $C_s$ are possible. We just assume that the constants are such that the class has at least one compact support member $g_0$. In the construction of a subclass for Theorem 5.1, we let all $g_i$'s and $h_i$'s be translates of $g_0$, setting

$$g_i(x - c_i) = p_i\, g_0(x)\,,$$

$$h_i(x-d_i) = p_i\, g_0(x)\,,$$

for some constants $c_i$ , $d_i$ . See figure 5.2 below.

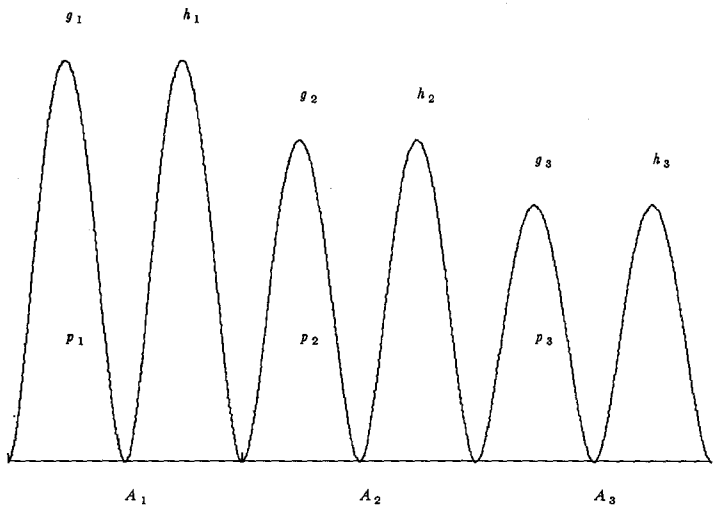


**Figure 5.2.**
Subclass of $\mathbf{F}$.

Note that every $f_\theta$ is in $\mathbf{F}$ because

$$g_i^{(s)}(x-c_i) = p_i\, g_0^{(s)}(x)\,.$$

**F consists of all monotone densities on $[0,\infty)$ that are bounded by 1.**

The lower bound of Theorem 5.1 is valid for this case too, but the proof needs some fine-tuning. Consider a set $A_1=[0,1]$, and an arbitrary number $p_1$. On $A_1$, define densities $g$ and $h$ as follows:

$$h(x) = 1 ,$$

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

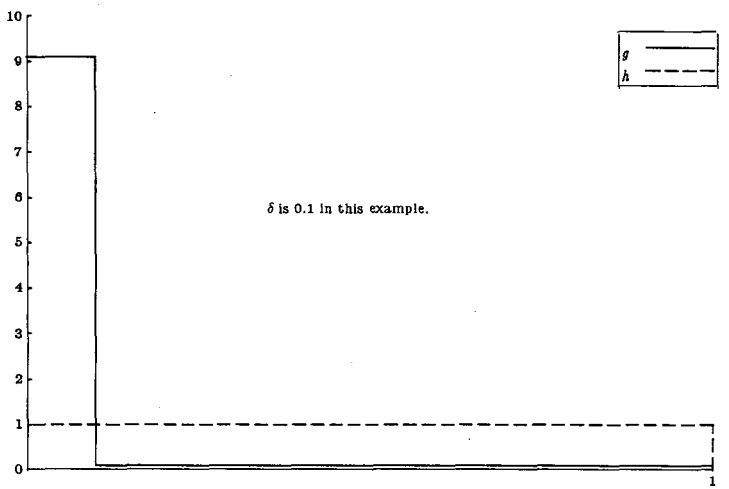


**Figure 5.3.**
Fundamental building block for $\int_\theta$.

Observe that $\int |g-h| = 2\int (h-g)_+ = 2(1-\delta)^2 \ge 2(1-2\delta)$. Partition $[0,\infty)$ into adjacent intervals, and define $g_i, h_i$ on each interval as $p_i$ times a properly translated and rescaled version of $g, h$. The translation is necessary because the intervals are nonoverlapping, and the rescaling is necessary to make the density monotone. We can assume without loss of generality that $p_1 \ge p_2 \ge p_3 \ge \cdots$. On every $A_i$, the functions $g_i, h_i$ are relatively positioned as in figure 5.3. See figure 5.4.

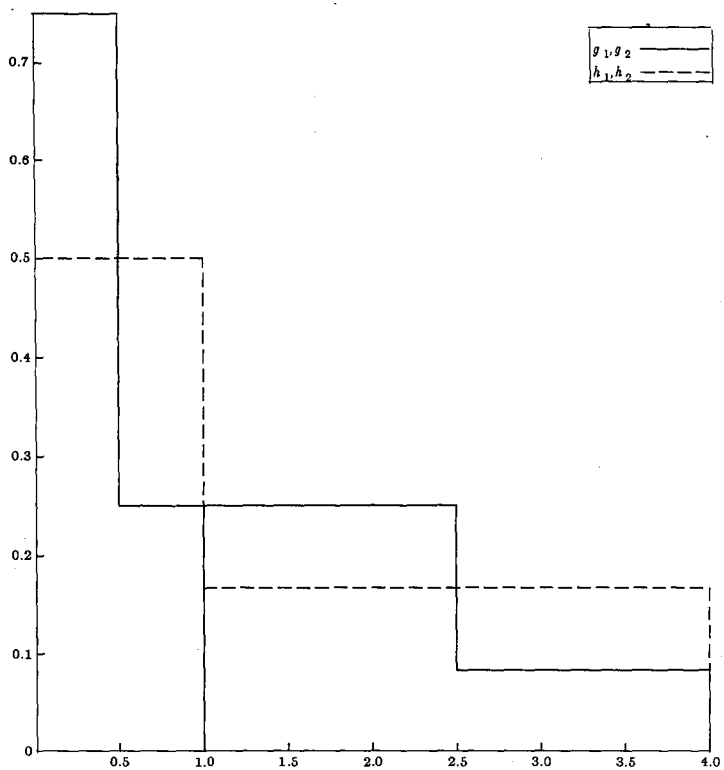


**Figure 5.4.**
Partition of real line, showing $g_1, g_2, h_1, h_2$ for $p_1 = p_2 = 0.5$.

For monotonicity, one should make sure that if $l_i$ is the length of the $i$-th interval $A_i$, then

$$\left(\frac{1}{\delta} - 1 + \delta\right)\frac{p_{i+1}}{l_{i+1}} = \delta\frac{p_i}{l_i}, \text{ all } i .$$

From this equation, and $l_1 = 1$, the lengths of the intervals can be determined in a recursive manner. The lower bound in the proof of Theorem 5.1 should be replaced by

$$\sum_i p_i (1-p_i)^n (1-2\delta) ,$$

which can be made arbitrarily close to 1 by choice of the $p_i$'s and $\delta$. In other words, Theorem 5.1 remains valid.

We conclude from this partial list of very rich classes $\mathbf{F}$ that no rate of convergence results are possible under tail conditions alone (see first example), under smoothness conditions alone (see second example), and under monotonicity conditions alone (see third example). It is not difficult to see that the third example can be modified so that smoothness and monotonicity conditions alone again do not suffice to get minimax errors converging to zero. The same is true under any combination of tail conditions and monotonicity conditions (by playing on the infinite peak at the origin). In all cases, the culprit for

$$m(n, \mathbf{F}) \geq 1$$

is the fact that the space is so big that we can construct enough virtually non-overlapping densities that can be combined in a convex manner to create a gigantic subclass of $\mathbf{F}$. In other words, the "size" of the space of densities should be limited in some sense.

## 5.4. INFORMATION-THEORETIC METHODS.

We begin with

**Theorem 5.2. Assouad's theorem.**

Let $r \geq 1$ be an integer, and let $\mathbf{F}$ contain all $f_\theta$'s where $\theta = 0.\theta_1\theta_2 \cdots \theta_r$ can take $2^r$ possible values. If $\zeta_i$ agrees with $\theta$ except in the $i$-th bit, if $A_0, A_1, \ldots, A_r$ is a partition of $R^d$, and if

$$\int \sqrt{f_\theta f_{\zeta_i}} \geq \beta > 0 \quad \text{(for all } \theta\text{)},$$

$$\int_{A_i} |f_\theta - f_{\zeta_i}| \geq \alpha > 0 \quad \text{(for all } \theta\text{)},$$

then

$$\sup_{f \in \mathbf{F}} E\left(\int |f_n - f|\right) \geq \sup_\theta E\left(\int |f_n - f|\right) \geq \begin{cases} \dfrac{r\alpha}{2}\left(1 - \sqrt{2 - 2\beta^n}\right) \\ \dfrac{r\alpha}{4}\beta^{2n} \end{cases}.$$

Another lower bound is $\dfrac{1}{2} r\alpha\left(1 - \sqrt{2n(1-\beta)}\right)$.

**Proof of Theorem 5.2.**

$$\sup_\theta E\left(\int |f_n - f_\theta|\right) \geq E\left(\int |f_n - f_\Theta|\right)$$

($\Theta$ is uniform over its $2^r$ possible values)

$$= 2^{-r} \sum_\theta \int\int |f_n(x, \mathbf{x}_n) - f_\theta(x)| \, dx \prod f_\theta(x_j) \, d\mathbf{x}_n$$

$$(\mathbf{x}_n = (x_1, \ldots, x_n))$$

$$= 2^{-r} \sum_\theta \int \sum_{i=1}^{r} \int_{A_i} |f_n(x, \mathbf{x}_n) - f_\theta(x)| \, dx \prod f_\theta(x_j) \, d\mathbf{x}_n$$

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

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

($\theta_{i-(i+)}$ agrees with $\theta$ subject to $\theta_i = 0(1)$)

$$\geq 2^{-r} \sum_\theta \int \sum_{i=1}^{r} \frac{\alpha}{2} \min\left( \prod f_{\theta_{i+}}(x_j), \prod f_{\theta_{i-}}(x_j) \right) d\mathbf{x}_n$$

$$\geq \frac{r\alpha}{2} \inf_{\theta, i} \int \min\left( \prod f_{\theta_{i+}}(x_j), \prod f_{\theta_{i-}}(x_j) \right) d\mathbf{x}_n \, .$$

The last integral can be bounded from below by

$$\frac{1}{2}\left(\int\sqrt{\prod f_{\theta_{i,+}}(x_j)\prod f_{\theta_{i,-}}(x_j)}\,d\mathbf{x}_n\right)^2$$
$$=\frac{1}{2}\left(\prod\int\sqrt{f_{\theta_{i,+}}(x_j)\,f_{\theta_{i,-}}(x_j)}\,dx_j\right)^2$$
$$\geq\frac{1}{2}\beta^{2n}\ .$$

The integral can also be bounded as follows from below:

$$1-\frac{1}{2}L_1\left(\prod f_{\theta_{i,+}}(x_j)\,,\prod f_{\theta_{i,-}}(x_j)\right)\geq 1-H_2\left(\prod f_{\theta_{i,+}}(x_j)\,,\prod f_{\theta_{i,-}}(x_j)\right)$$
$$=1-\sqrt{2-2\int\sqrt{\prod f_{\theta_{i,+}}(x_j)\,,\prod f_{\theta_{i,-}}(x_j)}\,d\mathbf{x}_n}$$
$$\geq 1-\sqrt{2-2\beta^n}\ .\blacksquare$$

Assouad's theorem tells us that we should find a subclass of $2^r$ densities, which can be pictured as sitting at the vertices of a cube in $R^r$, such that the $L_1$ distance between all neighbors is at least $\alpha$, the $L_1$ distance between vertices $k$ edge-lengths apart is at least $k\alpha$, and the $H_2$ distance between neighbors is not too large (see the condition involving $\beta$):

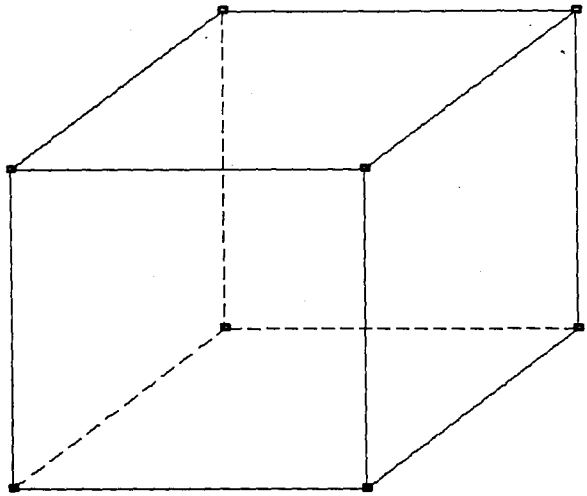

**Figure 5.5.**
Hypercube with $r = 3$ needed in Theorem 5.2.

Several examples of such constructions follow in the next few sections.

## 5.5. A CENTERED CLASS.

In a first example, we take the class of all densities that are within $(L_1)$ distance $\epsilon$ of a central density $f^*$. The example is analogous to an example worked out for Hellinger balls by Birge (1985).

**Theorem 5.3.**

$$\inf_{f_n} \sup_{f \in S(f^*,\epsilon)} E\left(\int |f_n - f|\right) \geq \epsilon .$$

We can interpret this theorem as follows. The estimator $f_n \equiv f^*$ is minimax-optimal for the class $\mathbf{F}$ since $\sup_{\mathbf{F}} \int |f^* - f| \leq \epsilon$. Thus, if minimizing the minimax criterion is our goal, then there is no need to consider the data; the data can in fact be thrown away, and we can take $f_n \equiv f^*$.

**Proof of Theorem 5.3.**

Partition the space (without loss of generality, we will take [0,1]) into $r$ sets of $f^*$-induced probability $1/r$. We will restrict $\mathbf{F}$ to all densities in $S(f^*,\epsilon)$ that have support [0,1]. On the $i$-th set $A_i$ in the partition, we define

$$h_i = \begin{cases} (1+\epsilon)f^* & \text{, between leftmost and median point of } A_i \\ (1-\epsilon)f^* & \text{, between median and rightmost point of } A_i \end{cases},$$

$$g_i = \begin{cases} (1-\epsilon)f^* & \text{, between leftmost and median point of } A_i \\ (1+\epsilon)f^* & \text{, between median and rightmost point of } A_i \end{cases}.$$

The median point in an interval is the point at which the integrals of $f^*$ over the two subintervals defined by the point are equal. See figure 5.6.

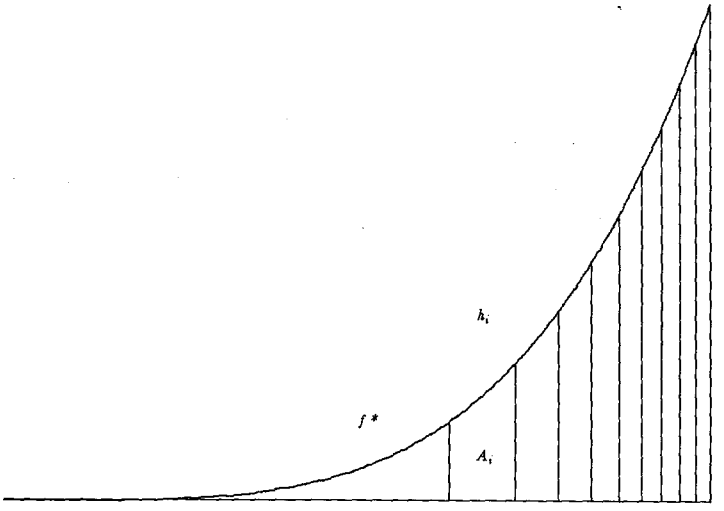


**Figure 5.6.**
Partition of $[0,1]$ into $r = 10$ equal probability segments.

Observe that

$$\int_{A_i} | h_i - g_i | = \frac{2\epsilon}{r} \stackrel{\Delta}{=} \alpha ,$$

$$\int | f_\theta - f^* | = \int \epsilon f^* = \epsilon \quad (\text{thus, all } f_\theta \in \mathbf{F}) ,$$

$$\int \sqrt{f_\theta f_{S_i}} = \frac{r-1}{r} + \frac{1}{r}\sqrt{1-\epsilon^2} \quad (\text{all } i)$$

$$= 1 - \frac{1}{r}(1-\sqrt{1-\epsilon^2}) \stackrel{\Delta}{=} \beta .$$

From Theorem 5.2, we obtain the lower bound

$$\frac{r\ \alpha}{2}\left(1-\sqrt{2-2\beta^n}\right) = \epsilon \left(1-\sqrt{2-2(1-\frac{1}{r}(1-\sqrt{1-\epsilon^2}))^n}\right)$$

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

$$\geq \epsilon \left(1-\sqrt{\frac{2n}{r}\epsilon^2}\right) ,$$

which is arbitrarily close to $\epsilon$ if $r$ is picked large enough. ■

## 5.6. A LIPSCHITZ CLASS.

For the first time in this chapter, we consider a truly small class, i.e. a class for which $m(n,\mathbf{F})$ decreases with $n$ as $n \to \infty$. Let $\mathbf{F}$ be the class of all Lipschitz densities on [0,1] with Lipschitz constant not exceeding $C$:

$$|f(x)-f(y)| \leq C|x-y| .$$

We also call this class $W(0,1,C)$, where "0" refers to the fact that the Lipschitz condition is for the 0-th derivative of $f$, "1" refers to the power of $|x-y|^1$ in the Lipschitz inequality, and $C$ is the constant. Not all constants $C$ are possible. In fact, $W(0,1,C)$ is not empty if and only if $C \geq 4$. We will prove the following

**Theorem 5.4.**

Let $C \geq 4$ be fixed and define $\rho = \sqrt{1-\frac{4}{C}}$. Then

$$\inf_{f_n} \sup_{f \in W(0,1,C)} E\left(\int |f_n - f|\right) \geq \left(\frac{162\,C}{10000\,n}\right)^{\frac{1}{3}} \sqrt{1-\frac{4}{C}} \left(1-\left(\frac{3C}{25\,n}\right)^{\frac{2}{3}}\right)^{\frac{1}{6}} - 1.22\,C^{-\frac{4}{3}} n^{-\frac{7}{6}} \rho^{-\frac{3}{2}} .$$

The inequality is valid for all $n$ and all $C \geq 4$. It is particularly useful to get a crude idea of the performance of any density estimate for a fixed value of $n$. Furthermore, for fixed $C$, the bound decreases as a constant times $(C/n)^{1/3}$. To better the bound by a factor of 10, $n$ should be increased by a factor of 1000. A sneak preview of things to follow: the kernel estimate will be shown to be minimax-optimal for $W(0,1,C)$ for particular choices of $K, h$.

**Proof of Theorem 5.4.**

A subclass of $\mathbf{F}$ is constructed by partitioning [0,1] into $r$ intervals of width $\frac{\rho}{r}$, where $\rho = \sqrt{1-\frac{4}{C}}$, and two intervals near the endpoints, accounting for a total length of $1-\rho$.

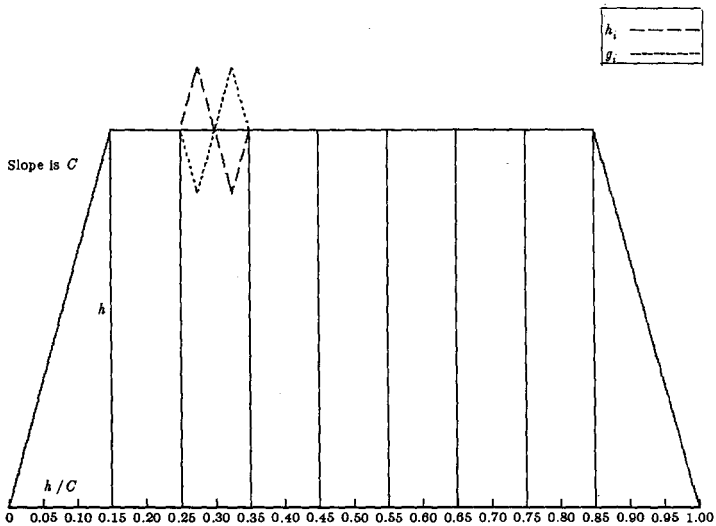


**Figure 5.7.**
Construction of a subclass for $W(0,1,C)$. Example with $r=7, \rho=0.7, C=7.84313725490$.

The two triangles near the endpoints are fixed: they do not vary with $n$. The slopes of the hypothenusas is $C$, and the height of the triangle is $h$ where

$$1-\frac{2h}{C} = \rho ,$$

i.e. $h = \frac{C}{2}(1-\rho)$. The value of $\rho$ can be determined from the requirement that the integral under the trapezoidal curve shown in figure 5.7 is one:

$$h\left(1-\frac{2h}{C}\right) + \frac{h^2}{C} = 1 .$$

This yields the value of $\rho$ suggested above. Every density in the subclass is equal to the central density with a small modification on each of the $r$ central intervals. The functions $h_i, g_i$ are as shown in the figure; the slopes of the curves are $\pm C$. Note that regardless of the value of $\theta$ (used in picking $g_i$ or $h_i$ in interval $A_i$), $f_\theta$ is a density. We also have

$$\int_{A_i} | h_i - g_i | = \frac{\rho}{2r} C \frac{\rho}{2r} = \frac{C\rho^2}{4r^2} \stackrel{\Delta}{=} \alpha .$$

and

$$\int \sqrt{f_\theta f_{\theta_i}} = 1 + 4 \int_0^{\frac{\rho}{4r}} \left( \sqrt{(h+Cx)(h-Cx)} - h \right) dx$$

$$= 1 - 4h \int_0^{\frac{\rho}{4r}} \left( 1 - \sqrt{1 - \frac{C^2x^2}{h^2}} \right) dx$$

$$\geq 1 - 4h \int_0^{\frac{\rho}{4r}} \frac{C^2x^2}{2h^2\sqrt{1-C^2x^2/h^2}} dx \quad \left(\text{since } \sqrt{1-u} \geq 1 - \frac{u}{2\sqrt{1-u}} \text{ for } 0 \leq u < 1\right)$$

$$\geq 1 - 4h \int_0^{\frac{\rho}{4r}} \frac{C^2x^2}{2h^2\phi(r)} dx$$

where

$$\phi(r) = \frac{1}{1 - \frac{C^2\rho^2}{16\, r^2h^2}} = \frac{1}{1 - \frac{\rho^2}{4\, r^2(1-\rho)^2}} .$$

Observe that $\phi(r)$ is decreasing in $r$. Taking the integral shows that

$$\int \sqrt{f_\theta f_{\theta_i}} \geq 1 - 2\, h\, \phi(r)\, \frac{C^2}{3h^2} \left( \frac{\rho}{4r} \right)^3$$

$$= 1 - \frac{\rho^3 C\, \phi(r)}{48\, r^3(1-\rho)} \stackrel{\Delta}{=} 1 - \gamma .$$

Assouad's lower bound now reads

$$\frac{r\,\alpha}{2} \left( 1 - \sqrt{2n\gamma} \right) = \frac{C\rho^2}{8\, r} \left( 1 - \sqrt{\frac{n\,\rho^3 C\, \phi(r)}{24\, r^3(1-\rho)}} \right)$$

$$\geq \frac{C\rho^2}{8\, r} \left( 1 - \sqrt{\frac{n\,\rho^3 C\, \phi(r_0)}{24\, r^3(1-\rho)}} \right) \quad (\text{all } r \geq r_0)$$

$$\stackrel{\Delta}{=} A \left( \frac{1}{r} - \frac{B}{r^{5/2}} \right) .$$

The right-hand-side of this expression should be maximized with respect to $r$. Setting the derivative with respect to $r$ equal to zero yields the value

$$r = \left(\frac{5B}{2}\right)^{\frac{2}{3}}.$$

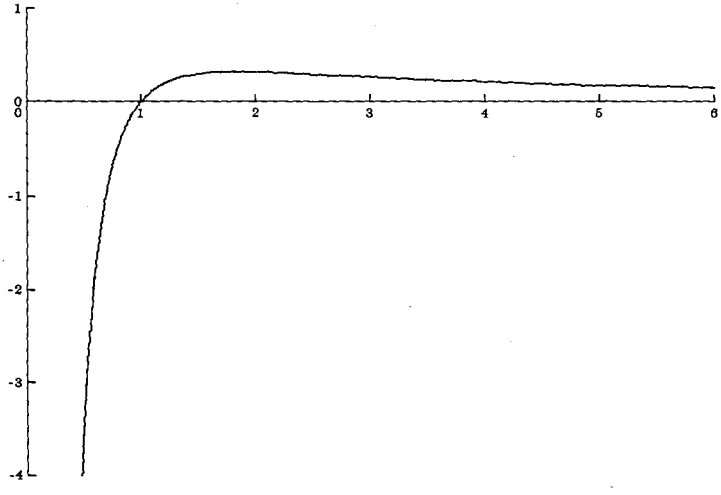


**Figure 5.8.**
The function $\frac{1}{r} - \frac{B}{r^{5/2}}$ for $B = 1$.

Unfortunately, $r$ has to be integer, a problem we will deal with further on. The maximal value is

$$\left(\frac{2}{5}\right)^{\frac{2}{3}} \frac{A}{B^{\frac{2}{3}}}\left(1-\frac{2}{5}\right) = \frac{3}{5}\left(\frac{2}{5}\right)^{\frac{2}{3}} \frac{A}{B^{\frac{2}{3}}}.$$

If we take

$$r_0 = \left(\frac{5B_0}{2}\right)^{\frac{2}{3}},$$

where $B_0$ is defined as $B$ with $\phi \equiv 1$, then $r \geq r_0$ in view of $B \geq B_0$. Resubstitution of the selected value of $r$ yields the bound

$$\frac{C\rho^2}{8} \frac{3}{5} \left(\frac{2}{5}\right)^{\frac{2}{3}} \left(\frac{24\,(1-\rho)}{n\,\rho^3 C\,\phi(r_0)}\right)^{\frac{1}{3}}$$

$$= \left(\frac{3}{10}\right)^{\frac{4}{3}} \frac{C^{\frac{2}{3}}(1-\rho)^{\frac{1}{3}}\rho}{n^{\frac{1}{3}}} \left(1-\frac{\rho^2}{4(1-\rho)^2 r_0{}^2}\right)^{\frac{1}{6}}.$$

Now, use the fact that

$$1-\rho \geq \frac{2}{C}$$

and that

$$(1-\rho)^2 r_0{}^2 = (1-\rho)^2 \left(\frac{5}{2}\right)^{\frac{4}{3}} \left(\frac{n\rho^3 C}{24(1-\rho)}\right)^{\frac{2}{3}} \geq \left(\frac{2}{C}\right)^{\frac{4}{3}} \left(\frac{25\, n\, \rho^3 C}{96}\right)^{\frac{2}{3}}$$

$$= \rho^2 \left(\frac{25\, n}{24\, C}\right)^{\frac{2}{3}}$$

to obtain the bound

$$\left(\frac{162\, C}{10000\, n}\right)^{\frac{1}{3}} \rho \left(1-\left(\frac{3\, C}{25\, n}\right)^{\frac{2}{3}}\right)^{\frac{1}{6}}.$$

This is the desired bound if we can make a final adjustment for the fact that $r$ is not integer. If we take

$$r = \lceil r_{\text{opt}} \rceil$$

where $r_{\text{opt}}$ is as before, then the maximal value obtainable for the lower bound as a function of $r$ is at least equal to the value at $r_{\text{opt}}$ (computed above) minus $AB/r_{\text{opt}}{}^5$. The correction factor is thus

$$\frac{AB}{r_{\text{opt}}{}^5} = \left(\frac{2}{5}\right)^{\frac{10}{3}} A\, B^{-\frac{7}{3}}$$

$$\leq 1.22\, C^{-\frac{4}{3}} n^{-\frac{7}{6}} \rho^{-\frac{3}{2}}$$

where we used the facts that $\phi(r_0) \geq 1, 1-\rho \leq 4/C$, and

$$\left(\frac{2^{37}3^7}{5^{20}}\right)^{\frac{1}{6}} = 1.21085\ldots . \blacksquare$$

For the class $W(s,\alpha,C)$, it is possible to obtain a bound of the following nature:

$$\inf_{f_n} \sup_{f \in W(s,\alpha,C)} E\left(\int |f_n - f|\right) \geq (c(s,\alpha)+o(1)) \left(\frac{C}{n^{s+\alpha}}\right)^{\frac{1}{1+2(s+\alpha)}}$$

where $c(s,\alpha)$ is a function of $s$ and $\alpha$ only. Observe that this provides us with a continuum of polynomial lower bounds with powers in the range $(-\frac{1}{2},0)$.

## 5.7. MIXTURE CLASSES.

Mixtures form important subclasses of densities. Consider for example the simple class $\mathbf{F}$ consisting of densities of the form

$$1 \pm \epsilon g(x-\frac{i}{r}) \quad (x \in A_i \stackrel{\Delta}{=} (\frac{i}{r}, \frac{i+1}{r})),$$

where $i=0,1,2,\ldots,r-1$, "1" is the uniform density on $[0,1]$, and $g$ is a fixed function on $[0,\frac{1}{r}]$ satisfying the following conditions: $|g| \leq 1$ (otherwise we wouldn't have nonnegative functions), and $\int g = 0$ (to assure that all functions integrate to one). The constant $\epsilon \in [0,1]$ is allowed to vary within $\mathbf{F}$. The $\pm$ gives us a choice for each of the $r$ intervals. Therefore, for each $\epsilon$, we have $2^r$ members in our family.

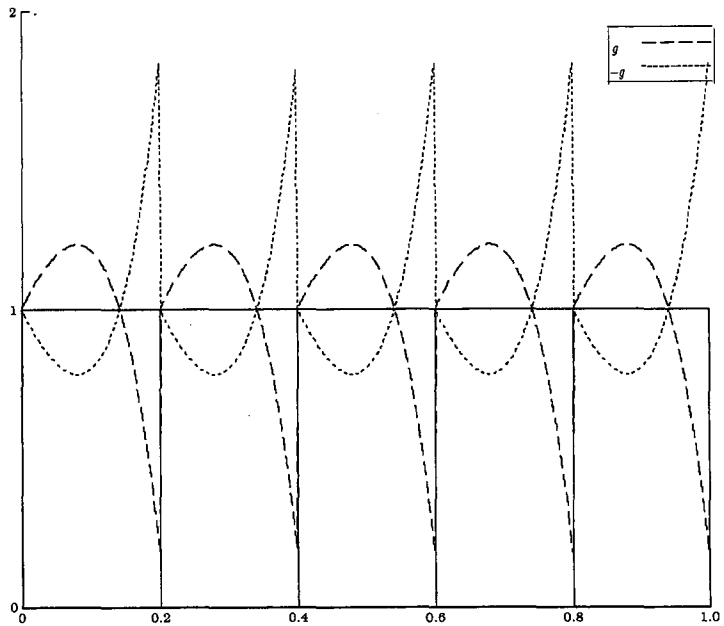


**Figure 5.9.**
Construction of a mixture class.

**Theorem 5.5.**

For the mixture class $\mathbf{F}$ defined on $r$ intervals,

$$\inf_{f_n} \sup_{f \in \mathbf{F}} E\left(\int |f_n - f|\right) \geq \sqrt{\frac{r}{32\,n}}, \quad n \geq \frac{r}{8}.$$

**Proof of Theorem 5.5.**

$$\int_{A_i} |h_i - g_i| = \int_0^{\frac{1}{r}} 2|g| \stackrel{\Delta}{=} \alpha,$$

$$\int \sqrt{f_\theta f_{\zeta_i}} = 1 - \int_0^{\frac{1}{r}} (1-\sqrt{1-g^2}) \stackrel{\Delta}{=} \beta.$$

Substitution of $\alpha$ and $\beta$ in Assouad's lower bound $(\frac{r\alpha}{2}(1-\sqrt{2n(1-\beta)}))$ with $|g| = \epsilon \leq 1$ yields the lower bound

$$\epsilon\left(1-\sqrt{\frac{2n}{r}(1-\sqrt{1-\epsilon^2})}\right)$$

$$\geq \epsilon - \epsilon^2\sqrt{\frac{2n}{r}}$$

$$= \sqrt{\frac{r}{32n}} \quad (\text{take } \epsilon = \sqrt{\frac{r}{8n}}).$$

The choice of $\epsilon$ maximizes the bound. The condition $\epsilon \leq 1$ implies that we should have $r \leq 8n$. ■

This simple bound illustrates clearly how doubling the number of intervals in the class can be balanced off by merely doubling the sample size $n$. Many other families have subfamilies of the form dealt with in Theorem 5.5, possibly even with $r$ increasing with $n$. In those cases, Theorem 5.5 provides a useful tool. See also exercises 5.2, 5.3.

## 5.8. CONVOLUTION CLASSES.

Assume that $\mathbf{F} = \{f_0 * \mu \mid$ all probability measures $\mu$ on the real line $\}$. Here $f_0$ is a fixed "central" density, and "*" is the convolution operator. An example of a related class is the class of all **normal scale mixtures**, i.e. the class of all densities of random variables $X$ distributed as $NY$ where $N$ is a normal (0,1) random variable and $Y$ is an arbitrary random variable. See the table below.

| $Y$ | $X=NY$ |
|---|---|
| 1/normal | Cauchy |
| $1/\sqrt{2}$ exponential | Laplace |
| $\sqrt{2a}$ /gamma$(a/2)$ | $t_a$ |

Observe that $\log|X|$ is distributed as $\log|N|+\log|Y|$ so that $\log|Y|$ defines the probability measure $\mu$ in the convolution, and the central density $f_0$ is the density of $\log|N|$. Convolution classes are harder to handle than the classes seen so far since member densities cannot conveniently be constructed on partitions of the space. Nevertheless, it is possible to prove

**Theorem 5.6.**

For any convolution class (i.e., any central density $f_0$),

$$\inf_{f_n} \sup_{f \in \mathbf{F}} E\left(\int |f_n - f|\right) \geq 1 .$$

In other words, meaningful minimax lower bounds are nonexistent if one looks at the class of all normal scale mixtures. Observe that the smoothness of each density in this class is determined by the smoothness of $f_0$, the normal (0,1) density, and that all members in the class are necessarily unimodal.

**Proof of Theorem 5.6.**

We will work with two parameters, a real number $\delta>0$, and an integer $r \geq 1$. First, it is necessary to find a number $M$ so large that

$$\int_{-M}^{M} f_0 > 1-\frac{\delta}{2} .$$

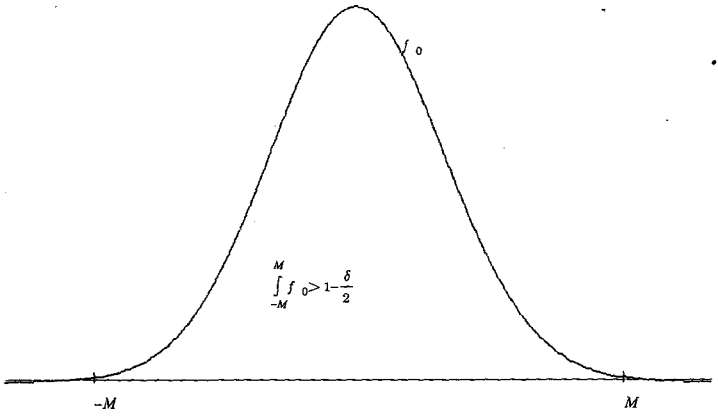


**Figure 5.10.**
Definition of $M$.

Consider next atomic measures $\mu_i$ with mass one atoms at points $x_i$, $i = -r, -r+1, \ldots, -1, 0, 1, \ldots, r-1, r$, where $x_i = 4iM$. The point will be to force the densities $f_0 * \mu_i$'s to be virtually non-overlapping. In a construction of a subfamily for Assouad's theorem, we define $\theta$ as in the theorem ($\theta$ has $r$ bits), and set

$$f_\theta = \frac{1}{r} \sum_{i=1}^{r} \left( f_0 * \mu_i I_{\theta_i = 1} + f_0 * \mu_{-i} I_{\theta_i = 0} \right) .$$

Define

$$A_i = [4iM - M, 4iM + M] \cup [-4iM - M, -4iM + M] .$$

Observe that

$$\begin{aligned} \int_{A_i} | f_\theta - f_{\zeta_i} | &\geq \frac{2}{r} \int_{4iM-M}^{4iM+M} \left( f_0 * \mu_i - f_0 * \mu_{-i} \right) \\ &\geq \frac{2}{r} \left( (1 - \frac{\delta}{2}) - \frac{\delta}{2} \right) \\ &= \frac{2}{r} (1 - \delta) . \end{aligned}$$

Also, using the notation

$$\bar{f}_i \stackrel{\Delta}{=} \frac{1}{2}\left(f_{\theta_{i+}}+f_{\theta_{i-}}\right),$$

$$\Delta_i \stackrel{\Delta}{=} \frac{1}{2}\left(f_{\theta_{i+}}-f_{\theta_{i-}}\right),$$

we have

$$\begin{aligned}
\int\sqrt{f_{\theta_{i+}}f_{\theta_{i-}}} &= 1-\int\left(\bar{f}_i-\sqrt{f_{\theta_{i+}}f_{\theta_{i-}}}\right)\\
&= 1-\int\left(\bar{f}_i-\sqrt{(f_i+\Delta_i)(f_i-\Delta_i)}\right)\\
&= 1-\int\bar{f}_i\left(1-\sqrt{1-\Delta_i^2/f_i^2}\right)\\
&\qquad (\Delta_i = (f_0*\mu_i - f_0*\mu_{-i})/2r)\\
&\geq 1-\int\bar{f}_i\frac{\Delta_i^2}{\bar{f}_i^2}\quad (\sqrt{1-u}\geq 1-u \text{ for } 0\leq u\leq 1)\\
&\geq 1-\int|\Delta_i| \quad (|\Delta_i|\leq\bar{f}_i)\\
&\geq 1-\frac{1}{r}.
\end{aligned}$$

Assouad's lower bound now becomes

$$\frac{r\left(\frac{2}{r}(1-\delta)\right)}{2}\left(1-\sqrt{\frac{2n}{r}}\right) = (1-\delta)\left(1-\sqrt{\frac{2n}{r}}\right).$$

The lower bound is arbitrarily close to one if we choose $\delta$ small enough and $r$ large enough. ■

## 5.9. FANO'S LEMMA.

Assouad's theorem is based upon a subclass of size $2^r$ of $\mathbf{F}$. The relationship between the member densities was illustrated with the help of a hypercube with $2^r$ vertices (see figure 5.5). In some cases, it is convenient to define subclasses of $r+1$ densities $f_\theta$, which can be viewed as vertices of a simplex in $r$-dimensional space (see figure 5.11):

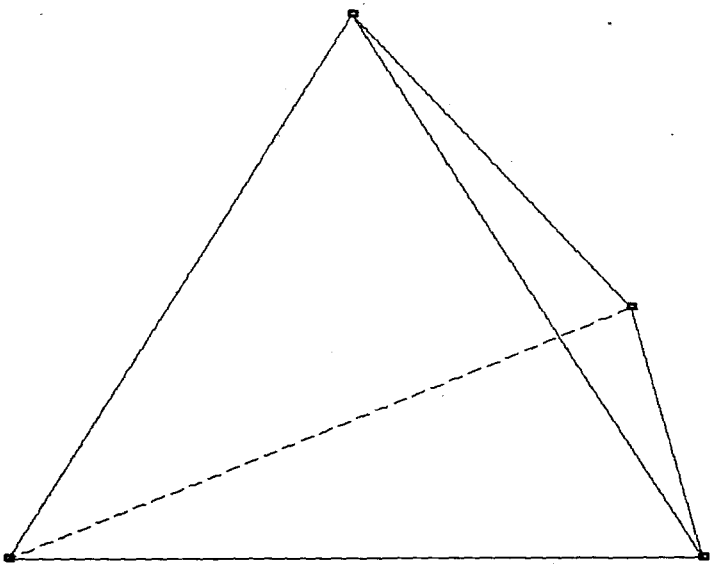

**Figure 5.11.**
Simplex with $r = 3$ needed in Theorem 5.7.

The $L_1$ distance between all densities must be at least equal to $\alpha$:

$$\inf_{\theta \neq \theta'} \int |f_\theta - f_{\theta'}| \geq \alpha ;$$

furthermore, the Kullback-Leibler distance between any pair of densities cannot be too large:

$$\sup_{\theta \neq \theta'} \int f_\theta \log\left(\frac{f_\theta}{f_{\theta'}}\right) \leq \beta .$$

The latter condition can be thought of as the counterpart of the upper bound on the Hellinger distances for the hypercube model. Assouad and Birge obtained the following generalization of Fano's lemma:

**Theorem 5.7. Generalization of Fano's lemma.**

Let $\mathbf{F}$ be a class of densities with a subclass of $r+1$ densities $f_\theta$ such that for any $\theta \neq \theta'$,

$$L_1(f_\theta, f_{\theta'}) \geq \alpha ,$$
$$K(f_\theta, f_{\theta'}) \leq \beta .$$

Then

$$\sup_{f \in \mathbf{F}} E(\int | f_n - f \; |) \geq \frac{\alpha}{2}\left(1 - \frac{n\;\beta + \log 2}{\log r}\right) .$$

We will proceed via two lemmas, including Fano's original lemma.

**Lemma 5.1.**

The **entropy** $-\sum_i p_i \log(p_i)$ of any probability vector $p_1, \ldots, p_n$ does not exceed $\log(n)$.

**Proof of Lemma 5.1.**

Let $\phi$ be a nonnegative convex function. Then

$$\frac{1}{n}\sum_i n \;\phi(p_i) \geq n\,\phi\left(\frac{1}{n}\sum_i p_i\right)$$

(Jensen's inequality)

$$= n\,\phi\left(\frac{1}{n}\right) .$$

Now, use the fact that $x \log(x)$ is convex in $x$. ■

**Lemma 5.2. Fano's lemma.**

Let $X$ be a random variable with density equal to one of $r+1$ possible densities $f_1, \ldots, f_{r+1}$, where $K(f_i, f_j) \leq \beta$ for all $i \neq j$. Let $\psi(X) \in \{1, \ldots, r+1\}$ be an estimate of the index. Then

$$\sup_i P_i(\psi(X) \neq i) \geq 1 - \frac{\beta + \log 2}{\log r},$$

where $P_i$ is the probability induced by $f_i$.

**Proof of Lemma 5.2.**

Let $\Theta$ be a random variable uniformly distributed on $1, \ldots, r+1$. Then

$$\begin{aligned}
&\sum_i P(\Theta = i \mid X) \log(P(\Theta = i \mid X)) \\
&= P(\Theta = \psi(X) \mid X) \log(P(\Theta = \psi(X) \mid X)) \\
&\quad + P(\Theta \neq \psi(X) \mid X) \log(P(\Theta \neq \psi(X) \mid X)) \\
&\quad + P(\Theta \neq \psi(X) \mid X) \sum_{i \neq \psi(X)} \frac{P(\Theta = i \mid X)}{P(\Theta \neq \psi(X) \mid X)} \log\left(\frac{P(\Theta = i \mid X)}{P(\Theta \neq \psi(X) \mid X)}\right) \\
&\geq -\log 2 - P(\Theta \neq \psi(X) \mid X) \log r,
\end{aligned}$$

where we applied Lemma 5.1 twice. The quantity on the left-hand-side of this chain of inequalities will now be bounded from above. Observe that

$$P(\Theta = i \mid X) = \frac{f_i(X)}{\sum_j f_j(X)}.$$

Thus,

$$\begin{aligned}
&E\left(\sum_i P(\Theta = i \mid X) \log(P(\Theta = i \mid X))\right) \\
&= \int \left(\sum_i \frac{f_i(x)}{\sum_j f_j(x)} \log\left(\frac{f_i(x)}{\sum_j f_j(x)}\right)\right) \frac{1}{r+1} \sum_j f_j(x)\, dx \\
&= \frac{1}{r+1} \sum_i \int \log\left(\frac{f_i(x)}{\sum_j f_j(x)}\right) f_i(x)\, dx \\
&\leq \frac{1}{(r+1)^2} \sum_{i,j} \int \log\left(\frac{f_i(x)}{f_j(x)}\right) f_i(x)\, dx - \log(r+1) \\
&\qquad \left(\text{use } \log\left(\frac{1}{r+1}\sum_j f_j\right) \geq \frac{1}{r+1}\sum_j \log(f_j)\right)
\end{aligned}$$

$$= \frac{1}{(r+1)^2}\sum_{i,j}\int K(f_i, f_j) - \log(r+1)$$
$$\leq \beta - \log(r+1) .$$

We conclude that

$$\beta - \log(r+1) \geq -\log 2 - P(\Theta \neq \psi(X))\log(r) .$$

Thus,

$$\sup_i P_i(\psi(X) \neq i) \geq P(\psi(X) \neq \Theta)$$
$$\geq \frac{\log(r+1) - \beta - \log 2}{\log r} ,$$

which was to be shown. ■

**Proof of Theorem 5.7.**

Let $X$ be the shorthand notation for $(X_1, \ldots, X_n)$, let $\Theta$ be as in the proof of Lemma 5.2, and let $g_n$ be defined as follows:

$$g_n = \begin{cases} f_\theta & \text{if } \int |f_n - f_\theta| < \alpha/2 \\ f_{\psi(X)} & \text{otherwise} \end{cases} .$$

Here $\psi(X)$ is defined arbitrarily (say, $\psi(X)=1$) in the "otherwise" case, and $\psi(X)$ is defined to be $\theta$ if $\int |f_n - f_\theta| < \alpha/2$. Observe that the $L_1$ balls of radius $\alpha/2$ centered at the functions $f_\theta$ do not overlap (by assumption). See figure 5.12.

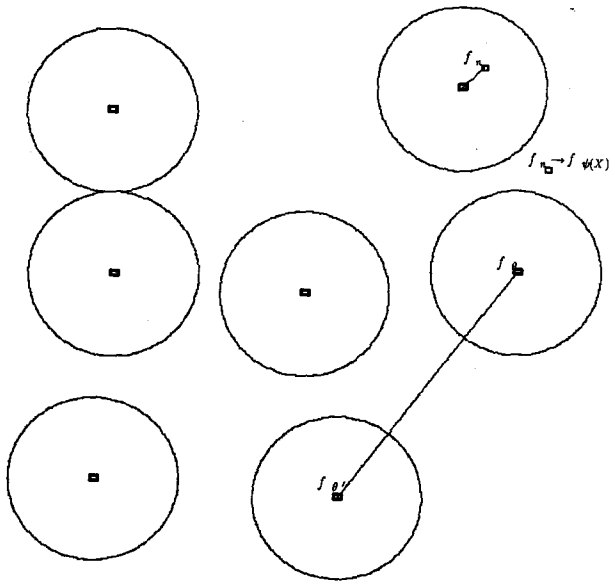

**Figure 5.12.**
Non-overlapping balls of radius $\alpha/2$ needed in the proof of Theorem 5.7. Distance between any two centers is at least $\alpha$.

The argument is very simple:

$$\max_\theta E_\theta(\int |f_n - f_\theta|) \geq \frac{\alpha}{2} \max_\theta P_\theta(\int |f_n - f_\theta| \geq \frac{\alpha}{2})$$

$$\geq \frac{\alpha}{2} \max_\theta P_\theta(\int |g_n - f_\theta| \geq \frac{\alpha}{2})$$

$$\geq \frac{\alpha}{2} \max_\theta P_\theta(\psi(X) \neq \theta)$$

$$\geq \frac{\alpha}{2}\left(1 - \frac{n\beta + \log 2}{\log r}\right)$$

by Lemma 5.2, and the fact that

$$K\left(\bigtimes_{k=1}^n f_i, \bigtimes_{k=1}^n f_j\right) = \int \log\left(\frac{\prod_{k=1}^n f_i(x_k)}{\prod_{k=1}^n f_j(x_k)}\right) \prod_{k=1}^n f_i(x_k) \prod_{k=1}^n dx_k$$

$$= n\ K(f_i, f_j) \leq n\ \beta\ ,$$

when $i \neq j$. ■

## 5.10. LOWER BOUNDS VIA SUFFICIENT STATISTICS.

There is yet a different method for obtaining lower bounds, which uses the fact that the best estimators depend upon sufficient statistics for the family only, and that the properties of simple sufficient statistics are often well-known. As an example, we take the class

$$\mathbf{F} = \{f \mid f = pg + (1-p)h\ ,\ p \in [0,1]\}$$

where $g$, $h$ are known densities with disjoint supports.

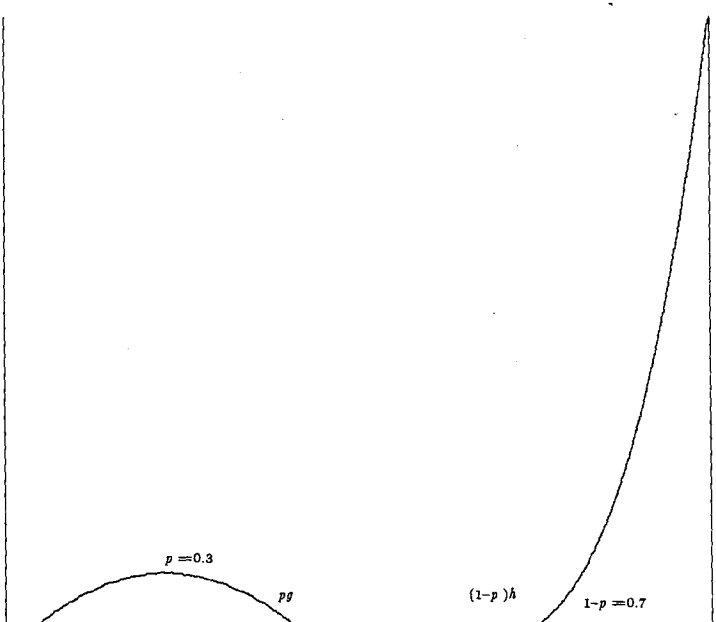


**Figure 5.13.**
A simple mixture.

**Theorem 5.8.**

Let $f_n$ be any density estimate with $\int f_n = 1$. Then

$$\sup_{f \in \mathbf{F}} E\left(\int |f_n - f|\right) \geq \begin{cases} \dfrac{0.030153}{\sqrt{n}}, & n \geq 4, \\ \dfrac{0.0849\ldots + o(1)}{\sqrt{n}} \end{cases}$$

**Proof of Theorem 5.8.**

Define $p_n = \int_A f_n$ where $A$ is the support of $g$. We first claim that

$$g_n = p_n\, g + (1-p_n)h$$

is at least as good an estimate as $f_n$. This follows from

$$\begin{aligned}\int | g_n - f | &= \int_A g\, | p_n - p | + \int_{A^c} h\, | p_n - p | \\ &= 2\, | p_n - p | = 2 \left| \int_A (f_n - f) \right| \\ &\le \int | f_n - f | \quad \text{(Scheffe's theorem)}.\end{aligned}$$

Next, we note that $N$, the cardinality of $A$, is a sufficient statistic for $p$. Thus, we should be able to reduce the data to $N$. This can be done by invoking the conditional form of Jensen's inequality:

$$\begin{aligned}E\left(\int | f_n - f |\right) &\ge 2\, E(| p_n - p |) \\ &\ge 2\, E(| E(p_n \mid N) - p |) \\ &= 2\, E(| \psi(N) - p |)\end{aligned}$$

where $\psi$ is some measurable function, which can be considered as an estimator based upon $N$ only. The dependence upon $n$ is suppressed temporarily. Observe that the joint density of the data given $N$ is

$$\frac{1}{n!} \sum_{\sigma} \prod_{i=1}^{N} f(x_{\sigma_i}) \prod_{j=N+1}^{n} g(x_{\sigma_j})$$

where $\sigma = (\sigma_1, \ldots, \sigma_n)$ is a permutation of $1, \ldots, n$. This density is independent of $p$. (This is another way of stating that in this problem, $N$ is a sufficient statistic for $p$.) This is why $E(p_n \mid N)$ is $\psi(N)$ and not $\psi_p(N)$. Now, we can randomize $p$, by making $p$ equal to $1/2$ with probability $1/2$, and equal to $1/2 + c/\sqrt{n}$ with probability $1/2$. Let $N_1, N_2$ be binomial random variables with parameters $(n, 1/2)$ and $(n, 1/2 + c/\sqrt{n})$ respectively. Then

$$\begin{aligned}&\sup_p E_p\left(\int | f_n - f |\right) \\ &\ge \frac{1}{2} E_{\frac{1}{2}}\left(\int | f_n - f |\right) + \frac{1}{2} E_{\frac{1}{2}+\frac{c}{\sqrt{n}}}\left(\int | f_n - f |\right) \\ &\ge E\left(\left| \psi(N_1) - \frac{1}{2} \right|\right) + E\left(\left| \psi(N_2) - \left(\frac{1}{2} + \frac{c}{\sqrt{n}}\right) \right|\right) \\ &\ge \sum_{j \le \frac{n}{2}} \left( \left| \psi(j) - \frac{1}{2} \right| P(N_1 = j) + \left| \psi(j) - \left(\frac{1}{2} + \frac{c}{\sqrt{n}}\right) \right| P(N_2 = j) \right) \\ &\ge \sum_{j \le \frac{n}{2}} \left( \left| \psi(j) - \frac{1}{2} \right| + \left| \psi(j) - \left(\frac{1}{2} + \frac{c}{\sqrt{n}}\right) \right| \right) P(N_2 = j)\end{aligned}$$

$$\geq \frac{c}{\sqrt{n}} P\left(N_2 \leq \frac{n}{2}\right).$$

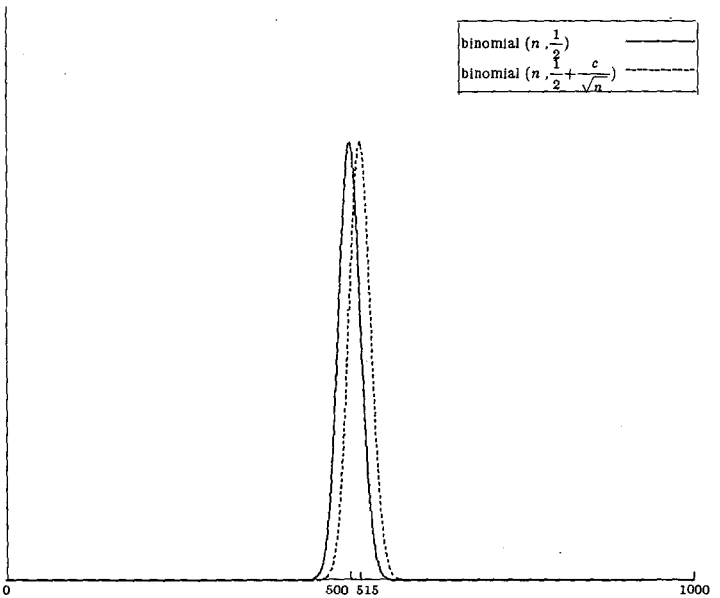


**Figure 5.14.**
Two smoothed binomial discrete densities. The parameters are $n = 1000$ and $c = 15/\sqrt{1000}$.

Now, let $Z$ be binomial $(n, \frac{1}{2}+\frac{c}{\sqrt{n}})$. Then

$$P\left(Z \leq \frac{n}{2}\right) = P\left(\frac{Z - E(Z)}{\sqrt{\mathrm{Var}(Z)}} \leq \frac{-c\sqrt{n}}{\sqrt{n\left(\frac{1}{2}+\frac{c}{\sqrt{n}}\right)\left(\frac{1}{2}-\frac{c}{\sqrt{n}}\right)}}\right)$$

$$\sim \Phi(-2c)$$

where $\Phi$ is the normal distribution function. Here we tacitly used the central limit theorem and a continuity argument. Thus, our lower bound is

$$\sim \frac{1}{\sqrt{n}} \left(c\,\Phi(-2c) + o(1)\right)$$

and has the largest possible value when

$$2c = 0.7517915241... \,. \blacksquare$$

## 5.11. CONSTRUCTION OF GOOD MINIMAX ESTIMATORS.

It would be cruel to develop many lower bounds and not to give examples of how one might find minimax-optimal estimators. Unfortunately, there seems to be no systematic approach at this point in time.

For very large classes $\mathbf{F}$, a suitably modified nonparametric estimate usually works well. A case in point is the kernel estimate for the Lipschitz classes $W(s,\alpha,C)$. More drastic modifications are needed for classes defined by order restrictions: this will be illustrated in another chapter for monotone densities.

Still for large classes $\mathbf{F}$, there are some attempts at making the construction systematic: in the next chapter, we will discuss minimum distance estimation (Yatracos, 1985); and the use of $\epsilon$-nets advocated by Birge (1986) can also be helpful.

Finally, for small $\mathbf{F}$, the nature of the classes can differ so dramatically that no general rules for constructing minimax-optimal estimators can be formulated. Each case needs to be handled separately.

## 5.12. EXERCISES.

5.1. Mimick the proof of Theorem 5.4 for the class $W(1,1,C)$, and obtain a lower bound with as main term a constant times $C^{1/5}n^{-2/5}$. Hint: replace the triangularly shaped $g_i, h_i$ in the construction of a subclass by quadratically shaped functions.

5.2. Improve the bound of Theorem 5.5 to $\frac{1+o(1)}{4}\sqrt{r/n}$, where $o(1)$ refers to asymptotics as $r/n \to 0$.

5.3. Consider the family of all densities of the form

$$f = \sum_{i=1}^{r} p_i g_i ,$$

where the $p_i$'s form a probability vector, and the $g_i$'s are (possibly overlapping) densities. The $g_i$'s are known, but the $p_i$'s are not. Derive a minimax lower bound for this class in terms of $n$, $r$ and possibly the distances between the $g_i$'s.

5.4. Consider the class $\mathbf{F}$ of all zero mean normal densities. Prove that

$$\inf_{f_n} \sup_{f \in \mathbf{F}} E\left(\int |f_n - f|\right) \geq \frac{c}{\sqrt{n}}$$

for some constant $c$. In a second step, construct a minimax-optimal estimate $f_n$.

5.5. Consider a family $\mathbf{F}$ satisfying the conditions of Assouad's theorem (Theorem 5.2), but with the upper bound on the Hellinger distance replaced by an upper bound on the Kullback-Leibler numbers

$$\int\int f_\theta \log\left(\frac{f_\theta}{f_{\zeta_i}}\right) \leq \beta$$

for all $\theta$ and all $i$. Show that

$$\inf_{f_n} \sup_{f \in \mathbf{F}} E\left(\int |f_n - f|\right) \geq \frac{\alpha}{4} e^{-\beta n} .$$

where $\alpha$ is as in Theorem 5.2 (Bretagnolle and Huber, 1979).

5.6. Consider the class $F_{s,C}$ of densities on [0,1] with the property that each member $f$ has $s-1$ absolutely continuous derivatives, and

$$D_s(f) \stackrel{\Delta}{=} \left(\int |f^{(s)}|\right)^{\frac{1}{2s+1}} \left(\int \sqrt{f}\right)^{\frac{2s}{2s+1}} \leq C < \infty .$$

Observe that $D_s(f)$ is a scale-invariant factor. Show that

$$\liminf_{n \to \infty} \inf_{f_n} n^{\frac{s}{2s+1}} \sup_{f \in F_{s,C}} E\left(\int |f_n - f|\right) \geq \gamma_s C$$

valid for all $C$ larger than some $C_s$, and some universal positive constant $\gamma_s$ depending upon $s$ only. Hints: define a subclass as in figure 5.7, with $h_i$ and $g_i$ replaced by smoother versions. Improperly scaled versions of a difference function $g$ can be obtained as follows: $g_2 = g_0 * g_1$ 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 $[-\frac{1}{2}, \frac{1}{2}]$. Note that $g_2^{(s)} = g_0^{(s-1)}(x+1/2) - g_0^{(s-1)}(x-1/2)$, so that $\int |g_2^{(s)}| = 2\int |g_0^{(s-1)}|$. Define $g = g_2(x+\frac{3}{4}) - g_2(x-\frac{3}{4})$ to force $\int g = 0$. Observe also that $g = 0$ outside $[-\frac{3}{2}, \frac{3}{2}]$. Construct $g_i$ as $f_0 + g$, and $h_i$ as $f_0 - g$ where $f_0$ is a central density, and $g$ is translated and rescaled to fit the partition model. Apply Theorem 5.2 or the inequality of the previous exercise (Bretagnolle and Huber, 1979).

5.7. Verify that the bound in Theorem 5.8 should be halved if the restriction that $\int f_n = 1$ is dropped.

5.8. Let $\mathbf{F}$ be the class of all convex densities on [0,1]. This class includes the J-shaped and U-shaped beta densities. Show that $m(n,\mathbf{F}) \geq c$ for some positive constant $c$ and all $n$.

5.9. Prove that for the class of all concave densities on [0,1], $m(n,\mathbf{F}) \geq cn^{-2/5}$ for some universal constant $c > 0$. Can this bound be improved?

5.10. Let $\mathbf{F}$ be the class of all densities whose characteristic function is zero outside $[-T,T]$. Show that $\liminf_{n\to\infty} m(n,\mathbf{F}) > 0$. Hint: to obtain a lower bound for the distance between two densities $f, g$ in the class, use the inequality $\int |f-g| \geq \sup_t |\phi-\psi|$, where $\phi,\psi$ are the characteristic functions corresponding to $f, g$.

# Chapter Six
# MINIMUM DISTANCE ESTIMATORS

## 6.1. DEFINITION.

We can now present a rather systematic, albeit computationally inefficient, method for constructing minimax-optimal estimates. Some conditions will have to be imposed on $\mathbf{F}$, so the construction is not universally applicable. However, the ideas are so interesting that it is difficult not to spend some time on minimum distance estimates.

The theory of minimum distance estimation for parameter estimation goes back to LeCam (1966). Other key papers include Pfanzagl (1968), Beran (1977), Pollard (1980), Millar (1981, 1983) and Yatracos (1985). Our treatment is mainly based on Yatracos (1985).

We recall that the **empirical measure** for $X_1, \ldots, X_n$ is defined by

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} I_{[X_i \in A]}$$

where $A$ is any Borel set. We do have one standing condition on $\mathbf{F}$, our class of densities: $\mathbf{F}$ is an $L_1$-totally bounded collection of densities, i.e. for every $\epsilon > 0$, $\mathbf{F}$ can be covered by a finite number of radius $\epsilon$ balls. In fact, it will be helpful to use special notation for such a cover: $\mathbf{F}_\epsilon$ is a finite collection of densities such that

$$\bigcup_{f \in \mathbf{F}_\epsilon} S(f, \epsilon)$$

covers $\mathbf{F}$. Sometimes is it useful to ask that $\mathbf{F}_\epsilon \subseteq \mathbf{F}$, but we will not impose this additional restriction.

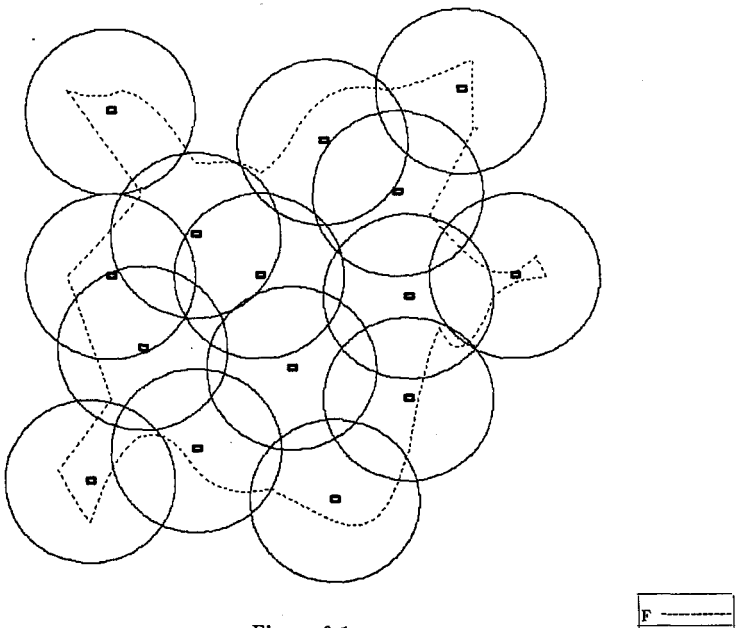


**Figure 6.1.**
$\mathbf{F}$ is covered by radius $\epsilon$ balls
centered at a finite number of densities from $\mathbf{F}$.

Let us also define a number that measures how "rich" $\mathbf{F}$ is,

$$N_\epsilon \stackrel{\Delta}{=} \inf_{\mathbf{F}_\epsilon} |\mathbf{F}_\epsilon| ,$$

where $|.|$ is the cardinality operator. Thus, $N_\epsilon$ is the size of the smallest $\epsilon$-cover of $\mathbf{F}$. The quantity $\log_2 N_\epsilon$ is also known as the **Kolmogorov entropy** of $\mathbf{F}$.

We are now ready for the definition of a minimum distance estimate. First choose $\epsilon$ according to the eventual expected error envisaged, and construct an $\epsilon$-cover of $\mathbf{F}$. The centers of the covering balls form a finite collection $\mathbf{F}_\epsilon$. Let $\mathbf{A}$ be the family of sets

$$\{x : f_\theta(x) > f_\theta(x)\} ,$$

where $f_\theta , f_\theta \in \mathbf{F}_\epsilon$. Note that $|\mathbf{A}| \leq |\mathbf{F}_\epsilon|^2$. We call $f_\theta$ the minimum distance estimate of $f$ if

$$f_\theta \in \mathbf{F}_\epsilon ,$$

and

$$2 \sup_{A \in \mathbf{A}} \left| \int_A f_\theta - \mu_n(A) \right| = \min_{f_\theta \in \mathbf{F}_\epsilon} \left( 2 \sup_{A \in \mathbf{A}} \left| \int_A f_\theta - \mu_n(A) \right| \right).$$

Observe that $\sup_{A \in \mathbf{A}} \left| \int_A f_\theta - \mu_n(A) \right|$ approximates the $L_1$ distance between $f_\theta$ and $\mu_n$. Obviously, had $\mathbf{A}$ be the collection of all Borel sets, then the distance would trivially have been 2 (as all discrete probability measures are at distance 2 from all absolutely continuous probability measures). On the other hand, if $\mathbf{A}$ is too small, there is no hope of extracting a reasonable $f_\theta$ from the collection. Hence we need a compromise on the size of $\mathbf{A}$. We should also keep in mind that the computation of $f_\theta$ requires time at least equal to $|\mathbf{A}|$ if no special computational shortcuts are used.

## 6.2. THE KEY INEQUALITY.

The purpose of this section is to prove

**Theorem 6.1.**

For $|\mathbf{A}| \geq 3$,

$$\sup_{f \in \mathbf{F}} E\left( \int |f_\theta - f| \right) \leq 5\epsilon + \frac{4}{\sqrt{2n}} \left( 1 + 2\sqrt{\log |\mathbf{A}|} \right).$$

The uniform upper bound depends upon $n$, $\epsilon$ and $|\mathbf{A}|$ only. In fact, if we construct our $\epsilon$-covers well, then the bound depends upon $\epsilon$, $n$ and the Kolmogorov entropy only. For a smallest possible bound, it is necessary to choose $\epsilon$ such that

$$5\epsilon + \frac{4}{\sqrt{2n}} \left( 1 + 2\sqrt{2 \log N_\epsilon} \right)$$

is minimal. Since $N_\epsilon$ increases and $\epsilon$ decreases as $\epsilon \downarrow 0$, the minimization problem is well-defined. We will see further on that $\epsilon$ is a function of $n$ and the class $\mathbf{F}$. Thus, this method tailors the estimates after the class $\mathbf{F}$ and $n$. This implies

that there could be an awful mismatch if $n$ is not as planned, or $f \notin \mathbf{F}$. In fact, this method is usually not consistent when $f \notin \mathbf{F}$.

The proof of Theorem 6.1 is in four stages:

**Lemma 6.1.**

For all $f, g \in \mathbf{F}$,

$$\int |f-g| \begin{cases} \leq 4\epsilon + 2 \sup_{A \in \mathbf{A}} \left| \int_A f - \int_A g \right| & \text{(Yatracos)} \\ \geq 2 \sup_{A \in \mathbf{A}} \left| \int_A f - \int_A g \right| & \text{(Scheffe)} \end{cases}$$

**Proof of Lemma 6.1.**

Find $\theta$ and $\theta'$ such that $f \in S(f_\theta, \epsilon)$ and $g \in S(f_{\theta'}, \epsilon)$. Then

$$\begin{aligned} \int |f-g| &\leq \int |f-f_\theta| + \int |f_\theta - f_{\theta'}| + \int |f_{\theta'} - g| \\ &\leq 2\epsilon + 2 \int_{f_\theta > f_{\theta'}} (f_\theta - f_{\theta'}) \\ &\leq 2\epsilon + 2 \int_{f_\theta > f_{\theta'}} (f_\theta - f) + 2 \int_{f_\theta > f_{\theta'}} (f-g) + 2 \int_{f_\theta > f_{\theta'}} (g - f_{\theta'}) \\ &\leq 4\epsilon + 2 \sup_{A \in \mathbf{A}} \left| \int_A f - \int_A g \right| . \blacksquare \end{aligned}$$

Lemma 6.1 establishes that the $L_1$ distance between two densities and the maximal deviation of induced probabilities differ by at most $4\epsilon$. Thus, there indeed is hope to obtain meaningful $L_1$ results by considering only sets from $\mathbf{A}$.

**Lemma 6.2. Gordon's inequality.**

For $t > 0$,

$$\int_t^\infty e^{-\frac{v^2}{2}}\, dv \leq \frac{1}{t} e^{-\frac{t^2}{2}} \quad \text{(Gordon's inequality)}.$$

This is a special case of a more general inequality, valid for functions $\phi \geq 0$, $\phi' \geq 0$, $\phi' \uparrow$, $\phi \uparrow \infty$:

$$\int_t^\infty e^{-\phi(v)}\, dv \leq \frac{e^{-\phi(t)}}{\phi'(t)}.$$

**Proof of Lemma 6.2.**

We will only prove the second part, as the first part is a simple corollary.

$$\begin{aligned}\int_t^\infty e^{-\phi(v)}\, dv &\leq \int_t^\infty \frac{\phi'(v)}{\phi'(t)} e^{-\phi(v)}\, dv \\ &= \frac{-1}{\phi'(t)} \int_t^\infty d\left(e^{-\phi(v)}\right) \\ &= \frac{e^{-\phi(t)}}{\phi'(t)}. \blacksquare\end{aligned}$$

**Lemma 6.3.**

For any probability measure $\mu$ and empirical measure $\mu_n$ (based upon an iid sample drawn from $\mu$), and for any collection of sets $\mathbf{A}$ with $|\mathbf{A}| \geq 3$,

$$E\left(\sup_{A \in \mathbf{A}} |\mu_n(A) - \mu(A)|\right) \leq \frac{2\sqrt{\log |\mathbf{A}|} + 1}{\sqrt{2n}}.$$

**Proof of Lemma 6.3.**

By Hoeffding's inequality (Bennett, 1962; Hoeffding, 1963), for any set $A$,

$$P(|\mu_n(A)-\mu(A)| \geq t) \leq 2e^{-2nt^2} \quad (t>0).$$

Thus,

$$E\left(\sup_{A\in\mathbf{A}} |\mu_n(A)-\mu(A)|\right)$$

$$\leq \int_0^\delta 2\,dt + \int_\delta^\infty 2|\mathbf{A}|e^{-2nt^2}\,dt$$

$$\text{(use } E|X| = \int_0^\infty P(|X|>t)\,dt \text{ and Bonferroni's inequality)}$$

$$= 2\delta + \frac{|\mathbf{A}|}{\sqrt{n}} \int_{2\delta\sqrt{n}}^\infty e^{-\frac{v^2}{2}}\,dv \quad (t = v/(2\sqrt{n}))$$

$$\leq 2\delta + \frac{|\mathbf{A}|}{\sqrt{n}} \frac{1}{2\delta\sqrt{n}} e^{-2n\delta^2} \text{ (Gordon's inequality)}$$

$$= 2\sqrt{\frac{\log|\mathbf{A}|}{2n}} + \frac{|\mathbf{A}|}{\sqrt{n}} \frac{1}{\sqrt{2\log|\mathbf{A}|}} \frac{1}{|\mathbf{A}|} \quad \left(\delta = \sqrt{\frac{\log|\mathbf{A}|}{2n}}\right)$$

$$\leq \frac{2\sqrt{\log|\mathbf{A}|}+1}{\sqrt{2n}}, (|\mathbf{A}| \geq 3).$$

The choice of $\delta$ in the last line is approximately optimal for large $|\mathbf{A}|$. Indeed, the derivative of

$$2\delta + \frac{|\mathbf{A}|}{2\delta n} e^{-2n\delta^2}$$

with respect to $\delta$ is

$$2 - \left(2+\frac{1}{2n\delta^2}\right)|\mathbf{A}|e^{-2n\delta^2} \approx 2 - 2|\mathbf{A}|e^{-2n\delta^2}$$

which is zero for the given choice of $\delta$ (note that for this choice, $2n\delta^2 \to \infty$ as $|\mathbf{A}| \to \infty$.). ■

**Proof of Theorem 6.1.**

$$\int |f_\theta - f| \leq 4\epsilon + 2\sup_{A\in\mathbf{A}} \left|\int_A f_\theta - \int_A f\right| \quad \text{(Lemma 6.1)}$$

$$\leq 4\epsilon + 2\sup_{A\in\mathbf{A}} \left|\int_A f_\theta - \mu_n(A)\right| + 2\sup_{A\in\mathbf{A}} \left|\mu_n(A) - \int_A f\right|$$

$$\leq 4\epsilon + 2\sup_{A\in\mathbf{A}} \left|\int_A f_\theta - \mu_n(A)\right| + 2\sup_{A\in\mathbf{A}} \left|\mu_n(A) - \int_A f\right|$$

$$\text{(for some } f_\theta \in S(f,\epsilon)\text{)}$$

$$\leq 4\epsilon + 2 \sup_{A \in \mathbf{A}} \left| \int_A f - \int_A f \right| + 4 \sup_{A \in \mathbf{A}} \left| \mu_n(A) - \int_A f \right| .$$

The expected value of the last expression does not exceed

$$5\epsilon + \frac{4}{\sqrt{2n}} \left(1 + 2\sqrt{\log |\mathbf{A}|}\right) \text{ (Lemma 6.3)} . \blacksquare$$

## 6.3. CONSTRUCTION OF AN $\epsilon$-COVER.

Let us explicitly construct an $\epsilon$-cover for the Lipschitz class $W(0,1,C)$, which in this section will simply be called the **Lipschitz class**. Consider first a grid as in figure 6.2. The interval $[0,1]$ is divided into $1/\delta$ equal intervals, where $1/\delta$ is an integer to be picked further on. For technical reasons, we also assume that $1/(C\delta)$ and $C$ are integer-valued. The $y$-axis is cut up into equal intervals of length $C\delta$. Consider as $\mathbf{F}_\epsilon$ all grid functions (functions that are allowed to follow the grid lines only are called grid functions), taking the value zero at $x=0$ and $x=1$, and moving up one level, or down one level, or not moving at all at each mesh point (see figure 6.2). In addition, the area under each $f_\theta$ must be one (since the area is a multiple of $C\delta^2$, this is feasible only if $1/(C\delta^2)$ is integer-valued).

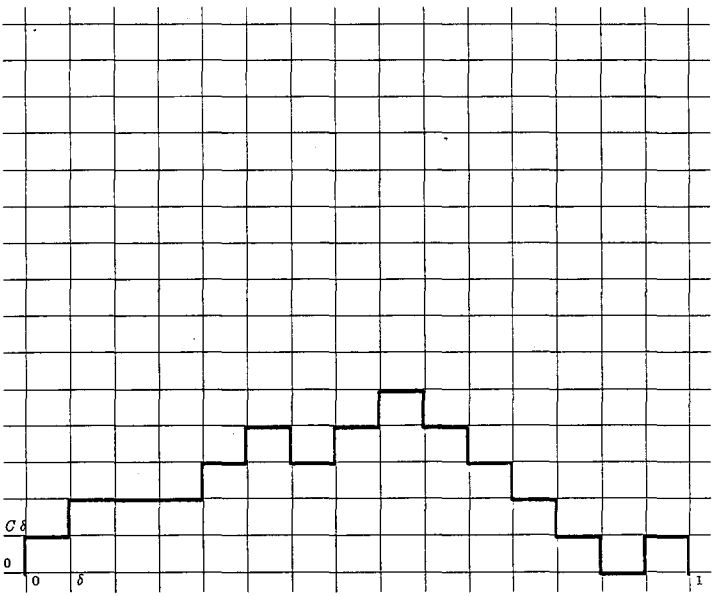


**Figure 6.2.**
Grid for the construction of $\mathbf{F}_\epsilon$.
A typical $f_\theta$ is shown.
$f_\theta$ takes steps $0, \pm 1$ from strip to strip.

One can easily verify that

$$|\mathbf{F}_\epsilon| \leq 3^{\frac{1}{\delta}}$$

because at each mesh point, we can only make one of three decisions. Furthermore,

$$|\mathbf{A}| \leq 2^{\frac{1}{\delta}}$$

In view of the fact that there are only $2^{\frac{1}{\delta}}$ possible subsets of $\frac{1}{\delta}$ elements. For a function $f \in \mathbf{F}$, find the values $f(i\delta)$ for all integer $i$, and round these to the nearest mesh point value (multiple of $C\delta$). Then consider a grid function $g$ (not necessarily with the correct area) passing through these points (see figure 6.3).

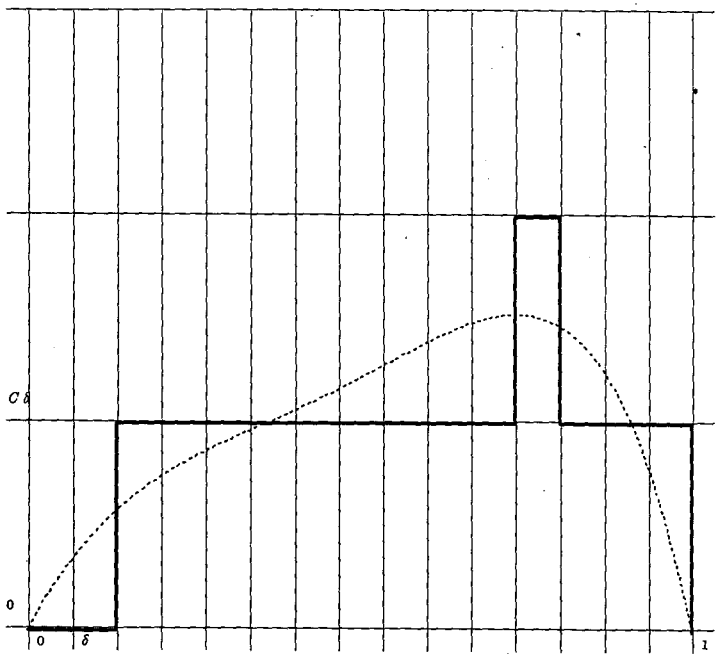


**Figure 6.3.**
A Lipschitz density with $C=15$, and its grid approximation $g$.

For this function $g$, we compute the area, and obtain a grid function $f_\theta \in \mathbf{F}_\epsilon$ with the property that

$$\int |f_\theta - g| = |1 - \int g|$$

(this is always possible). We have

$$\int |g - f_\theta| \leq \frac{1}{2\delta}\frac{1}{2}\left(\frac{1}{2}C\delta^2 + \frac{3}{2}C\delta^2\right) = C\delta.$$

Furthermore, since $|\int g - 1| \leq C\delta$, we have $\int |f_\theta - f_\theta| \leq 2C\delta \leq \epsilon$, if we take

$$\frac{1}{C\delta} = \left\lceil \frac{2}{\epsilon} \right\rceil.$$

Hence, we have an $\epsilon$-cover of $\mathbf{F}$. Observe that $\log |\mathbf{A}| \leq \log_2 |\mathbf{A}| \leq 1/\delta$, so that by Theorem 6.1,

$$\sup_{f \in \mathbf{F}} E\left(\int |f_\theta - f|\right) \le 5\epsilon + \frac{4}{\sqrt{2n}}\left(1+\frac{2}{\sqrt{\delta}}\right)$$

$$\le 5\epsilon + \frac{4}{\sqrt{2n}} + \frac{8\sqrt{2C}}{\sqrt{n\,\epsilon}} \quad \left(\frac{1}{\delta} \le \frac{4C}{\epsilon} \text{ for } \epsilon \le 2\right)$$

$$= \frac{4}{\sqrt{2n}} + 3\,4^{\frac{2}{3}}\,5^{\frac{1}{3}}\left(\frac{2C}{n}\right)^{\frac{1}{3}} \quad \text{(optimal choice of } \epsilon\text{)}.$$

The optimal choice of $\epsilon$, obtained by setting the derivative with respect to $\epsilon$ equal to zero, is

$$\epsilon = \left(\frac{4}{5}\right)^{\frac{2}{3}}\left(\frac{2C}{n}\right)^{\frac{1}{3}}.$$

For this to be smaller than 2 (a condition used in the chain of inequalities), we need to require that $n \ge 4C/25$.

It should be noted that the upper bound matches a lower bound obtained in chapter 5 up to a constant. In other words, the minimum distance estimate is minimax-optimal. There is a price tag: the computation time grows roughly speaking as

$$2^{\frac{1}{\delta}} \approx 2^{cn^{\frac{1}{3}}}$$

for some constant $c$ depending upon $C$ only. When $C$ is not integer, replace $C$ by $\lceil C \rceil$ in the upper bound, and observe that in any case, $\lceil C \rceil \le 5C/4$.

## 6.4. KOLMOGOROV'S ENTROPY.

Kolmogorov and Tikhomirov (1959) and Clements (1963) have shown that for $W(s,\alpha,C)$ in $R^d$,

$$N_\epsilon \approx 2^{\left(\frac{1}{\epsilon}\right)^{\frac{d}{s+\alpha}}}.$$

For $W(s,\alpha,C)$ and $d=1$, it is possible to construct a collection $\mathbf{F}_\epsilon$ with $|\mathbf{A}| \le N_\epsilon^2$. The corresponding minimum distance estimate satisfies

$$\sup_{f \in \mathbf{F}} E\left(\int |f_\theta - f|\right) \le 5\epsilon + \frac{4}{\sqrt{2n}} + \sqrt{\frac{16 \log N_\epsilon}{n}},$$

which, in view of the asymptotic expression for $N_\epsilon$ (valid as $\epsilon \downarrow 0$) forces us to minimize

$$5\epsilon + \frac{c}{\sqrt{n}}\epsilon^{-\frac{1}{2(s+\alpha)}}$$

where $c = \sqrt{16 \log 2}$. This is minimal for

$$\epsilon = \left( \frac{c^2}{100\,(s+\alpha)^2\, n} \right)^{\frac{s+\alpha}{1+2(s+\alpha)}}.$$

Resubstitution in the original bound gives us the result

$$\sup_{f \in F} E(\int |f_{\hat{}} - f|) = O\left( n^{-\frac{s+\alpha}{1+2(s+\alpha)}} \right).$$

Observe that this matches lower bounds obtained via Assouad's theorem (Theorem 5.2), at least when one only considers the dependence upon $n$.

The Lipschitz classes $W(s,\alpha,C)$ are perhaps the most widely studied totally bounded classes. Another example is the class of all nonincreasing densities on [0,1] bounded by a constant $c$. But some rather small classes are not totally bounded, such as the translation class

$$\{f(x-a) \mid a \in R\}$$

or the scale class

$$\{\frac{1}{\sigma} f(\frac{x}{\sigma}) \mid \sigma > 0\}$$

where $f$ is a fixed density. Examples include the normal densities with variable mean and/or variance.

## 6.5. EXERCISES.

6.1. Give an explicit construction of an $L_1$ minimum distance estimator, complete with a good upper bound for $|\mathbf{A}|$ and explicit expressions for $\epsilon$ and other parameters as a function of $C, \alpha, n$, for the class $W(0,\alpha,C)$ where $\alpha \in (0,1]$. The construction of section 6.3 for $\alpha = 1$ should be a special case of your construction.

6.2. Let $\mathbf{F}$ be the class of densities on [0,1] with modulus of continuity not exceeding $\omega(\delta)$. Show that there exist positive constants $c_1, c_2$ such that

$$\log(N_\epsilon) \leq \frac{c_1}{\omega^{-1}(c_2 \epsilon)}.$$

(Lorentz, 1966).

# Chapter Seven
# RATE OF CONVERGENCE OF KERNEL ESTIMATES

## 7.1. SCOPE OF THIS CHAPTER.

The purpose of this chapter is to give some idea of the relationship between the smoothness of a density and the best possible rates of convergence that can be attained by the kernel estimate. The story also depends upon the choice of $K$: for example, it does matter whether $K \geq 0$ or not.

The classes of densities we are dealing with are very large. Although there are other estimates that can provide good rates of convergence for these classes, the kernel estimate certainly is the most important one from a didactic point of view.

We will study two quantities, **the uniform performance**

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

and **the individual performance**

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

Unfortunately, it is once again necessary to limit our treatment somewhat. In both cases, we will study small sample and asymptotic upper bounds. In the case of the uniform performance, that is not a major concession, since the upper bounds usually match minimax lower bounds for the given classes up to a small constant (in other words, the kernel estimate is minimax-optimal). It is unfortunate that we won't be able to cover the individual performance in detail. Lower bounds for $E\left(\int |f_n - f|\right)$ provide us with information about how large $n$ should be for any given density. For example, a result for $d=1$ not covered here is

$$\inf_{f,K} \liminf_{n \to \infty} \inf_{h} n^{\frac{2}{5}} E\left(\int |f_n - f|\right) \geq 0.86 .$$

Here the infimum is over all $f$, all smoothing factors $h$ and all nonnegative kernels $K$ (Devroye and Penrod, 1984). This result roughly states that we are bound to make $L_1$ errors at least equal to $0.86\, n^{-2/5}$. To have errors of the order of 0.01, we would need $n$ of the order of 100000 or bigger, even for the nicest $f$. For practitioners, this is a sobering result. It is possible to do better if special estimates are used for specific classes of densities; these estimates could be disastrous if used on densities outside the designated classes, but at least we can get somewhere with moderate sample sizes for some $f$.

Finally, we should stress that we are not concerned for the time being with methods of choosing $h$ as a function of the data. In this chapter, $h$ is allowed to depend upon $n$ only.

## 7.2. CLASSES OF KERNELS.

The main class of kernels of interest to us is the class of symmetric (about 0) functions integrating to one. These are called the **class 0** kernels. They include all symmetric densities. A **class s** kernel is a class 0 kernel for which

$$\int |x|^s |K(x)|\, dx < \infty,$$

and

$$\int x^i K(x)\, dx = 0$$

for all $i = 1,\ldots,s-1$, where $s$ is a positive integer. Thus, most class 0 kernels are in fact class 2 kernels, the only additional condition being that $|x|^2 K(x)$ have a finite integral. However, nonnegative class 0 kernels cannot possibly be class $s$ kernels for $s \geq 3$. In view of the symmetry of all kernels considered here, we need only discuss even values of $s$.

The **order** of a class 0 kernel is the largest integer $s$ such that $K$ belongs to class $s$. If $K$ is in class $s$ for all $s$, then $K$ is called a **superkernel.**

Let us give a few examples before moving on to other kinds of kernels. The kernel

$$K(x) = \frac{3}{4}(1-x^2)_+$$

is nonnegative and symmetric, and integrates to one. Its order is 2 since it has compact support. However, the Cauchy density has order 0, since its first absolute moment is infinite. If we want to construct order 4 kernels of compact support, we have to consider negative-valued functions.

Class $2s$ kernels can be constructed in a systematic manner in a number of ways. For example, $K$ can be fit into a symmetric polynomial model on a compact set with $s$ unknowns. There are $s$ conditions to be satisfied, which yield $s$ linear inequalities with $s$ unknowns. For example, if we assume that

$$K(x) = a + bx^2$$

on [-1,1], then $K$ is in class 4 if

$$a + \frac{b}{3} = \frac{1}{2},$$

$$\frac{a}{3} + \frac{b}{5} = 0.$$

The solution is $K(x) = \frac{1}{8}\left(9-15x^2\right)$ $(|x| \leq 1)$.

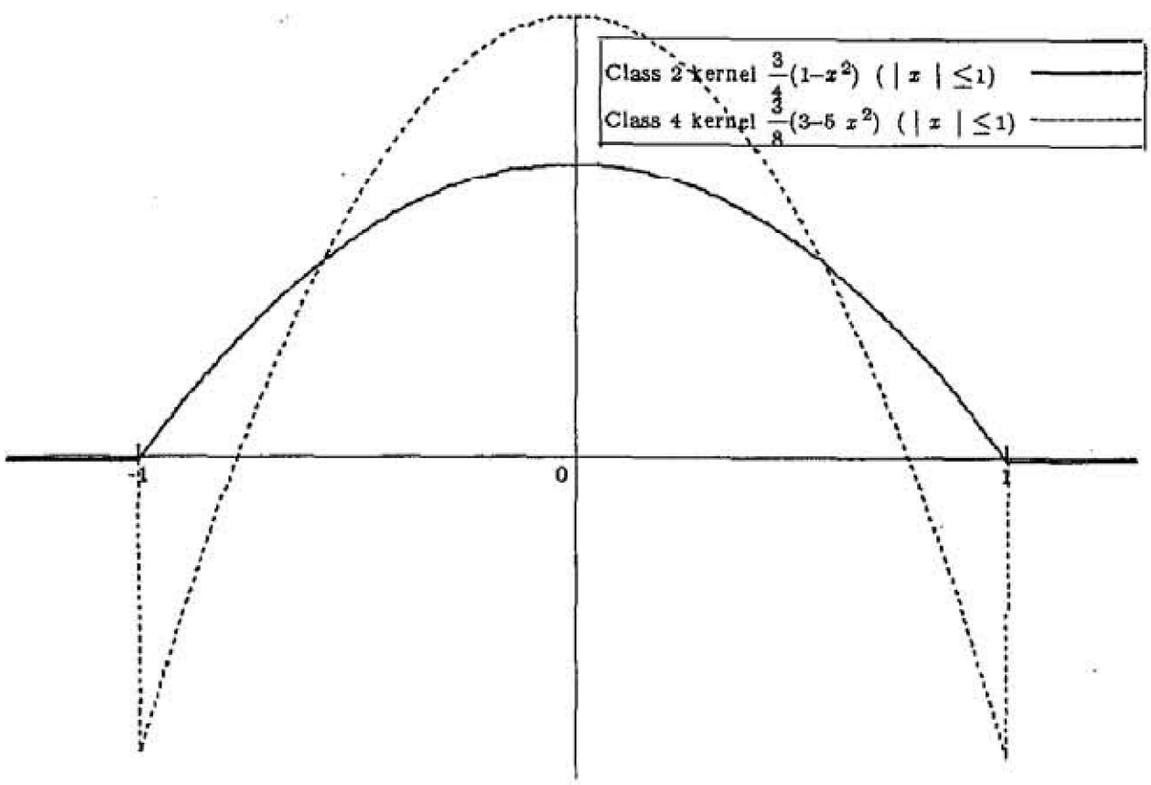


**Figure 7.1.**
A class 2 kernel $(K(x) = \frac{3}{4}(1-x^2)_+)$, and a class 4 kernel.

For the construction of superkernels, one needs to devise another method since we would need to solve an infinite number of equations with an infinite number of unknowns. A completeness argument shows that compact support superkernels do not exist since $\int_0^1 x^n f(x)\,dx = 0$ for all $n \geq 0$, $f \in L_1(0,1)$ together imply $f \equiv 0$ on (0,1) (see e.g. Theorem 22 of Hardy and Rogosinski (1962)). In fact, this conclusion remains valid if the condition were to hold for a subsequence of integers $n_1 < n_2 < \cdots$ for which $\sum_i n_i^{-1} = \infty$ (Muntz's theorem, see e.g. Steinhaus and Kaczmarz, 1935). We recall here that if $K$ has characteristic function $\psi$ (i.e., $\psi(t) = \int e^{itx} K(x)\,dx$), then $\psi(0) = 1$ for class 0 kernels,

and $\psi$ is real and even by the symmetry of $K$. Also, for even $s$, $(-1)^{s/2}\psi^{(s)}(0)=\int x^s K(x)dx$. Thus, for class $s$ kernels, with $s$ even, it suffices to invert a real even $\psi$ with $\psi(0)=1$ and $\psi^{(i)}(0)=0$ for all even $i \leq s$. To illustrate this, we start with a function $\psi$ with all zero derivatives at the origin, such as

$$\psi(t) = 1-e^{-\frac{1}{t^2}},$$

or

$$\psi(t) = (2-|t|)_+ - (1-|t|)_+ = \begin{cases} 1 \text{ on } [-1,1] \\ 2-|t| \text{ on } [-2,-1]\cup[1,2] \\ 0 \text{ outside } [-2,2] \end{cases}.$$

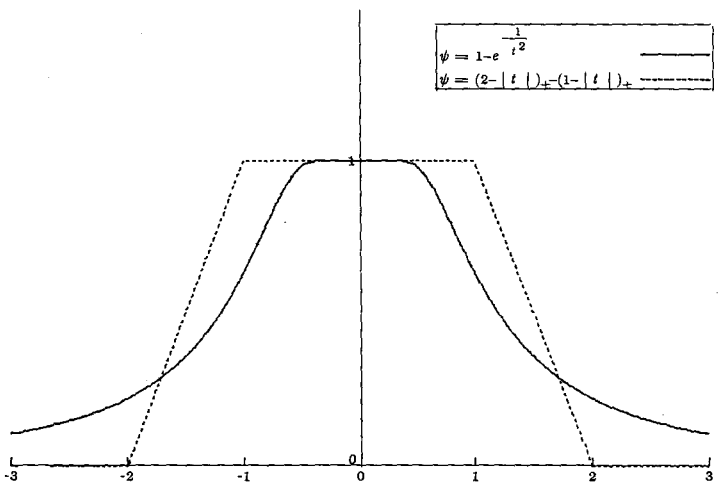


**Figure 7.2.**
Two characteristic functions used to construct superkernels.

Since $\int |\psi| < \infty$ in both cases, the following inversion formula is valid:

$$K(x) = \frac{1}{2\pi}\int \cos(tx)\,\psi(t)\,dt\ .$$

The kernel $K$ can be a superkernel. The second $\psi$ shown in figure 7.2 is flat in a neighborhood of the origin. The corresponding kernels will be referred to as **flattop kernels.** (Note that the kernels themselves do not have a flat part; the

characteristic function $\psi$ is 1 on an interval $[-T, T]$.) If we recall that $(1-|t|)_+$ is the characteristic function of the **de la Vallee Poussin density**

$$V(x) = \frac{1}{2\pi}\left(\frac{\sin(\frac{x}{2})}{\frac{x}{2}}\right)^2$$

shown in figure 7.3, then a small argument shows that the characteristic function $(2-|t|)_+ - (1-|t|)_+$ corresponds to the kernel

$$K(x) = 4V(2x) - V(x) = \frac{1}{2\pi x^2}\left(4\sin^2(x) - 4\sin^2(\frac{x}{2})\right)$$

$$= \frac{1}{2\pi x^2}\left(2\cos(x) - 2\cos(2x)\right) = \frac{2}{\pi x^2}\sin(\frac{x}{2})\sin(\frac{3x}{2}).$$

This kernel is also shown in figure 7.3. A simple computation shows that

$$\int |K| \leq 3, \sup_x |K(x)| \leq \frac{\int |\psi|}{2\pi} = \frac{3}{2\pi}.$$

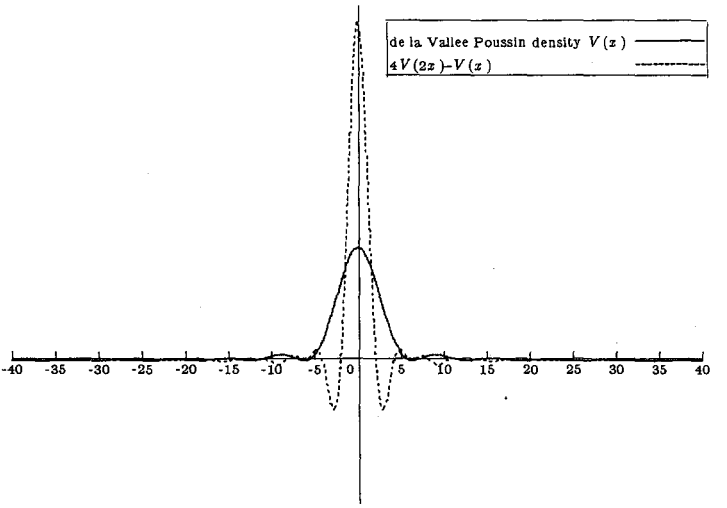


**Figure 7.3.**
de la Vallee Poussin density, and corresponding flattop kernel.

Observe that $\int x^2 |K(x)|\, dx = \infty$, so that strictly speaking $K$ is not in class 2

(even though $\int x^i K(x)\,dx = 0$ for all positive $i$). To force $K$ to have an infinite order, it is necessary to consider $\psi$'s that are a lot smoother. The first characteristic function of figure 7.2 is so smooth that all derivatives exist, and are continuous and absolutely integrable. By the inequality

$$|K(x)| \leq \frac{\int |\psi^{(s)}|}{2\pi |x|^s}$$

valid for $s = 0,1,2,\ldots$, it is easily seen that $K$ is indeed a superkernel. See exercise 7.1.

We will also need the notion of an **associated kernel** $L$, as introduced by Bretagnolle and Huber (1979). The function $L$ defied by

$$L(x) = (-1)^s \int_x^\infty \frac{(y-x)^{s-1}}{(s-1)!} K(y)\,dy \quad (x>0)$$

$$L(-x) = (-1)^s L(x) \quad (x<0)$$

is the kernel associated with kernel $K$. Sometimes we will say that $L$ has parameter $s$, since this figures in the definition of $L$. When $K$ is symmetric, $L$ is symmetric. Furthermore,

$$\int |L| \leq \frac{1}{s!} \int |x|^s |K(x)|\,dx$$

for all nonnegative integers $s$. For $s=0$, we define $L=K$. For $K \geq 0$, we have the equality

$$\int |L| = \frac{1}{s!} \int |x|^s K(x)\,dx .$$

Finally,

$$\int L = \int \frac{x^s}{s!} K(x)\,dx = \begin{cases} = 0 , & s \text{ odd} \\ = 0 , & s \text{ even, and the order of } K \text{ is} > s \end{cases}$$

These statements are easily proved:

$$\int |L| = 2\int_0^\infty \left| \int_x^\infty \frac{(y-x)^{s-1}}{(s-1)!} K(y)\,dy \right| dx$$

$$\leq 2\int_0^\infty \left( \int_0^y \frac{(y-x)^{s-1}}{(s-1)!}\,dx \right) |K(y)|\,dy$$

$$= 2\int_0^\infty \frac{|y|^s}{s!} |K(y)|\,dy .$$

Here we have equality when $K \geq 0$. For even $s$, we have, similarly,

$$\int L = 2(-1)^s \int_0^\infty \left( \int_x^\infty \frac{(y-x)^{s-1}}{(s-1)!} K(y)\,dy \right) dx$$

$$= 2(-1)^s \int_0^\infty \left( \int_0^y \frac{(y-x)^{s-1}}{(s-1)!}\, dx \right) K(y)\, dy$$

$$= 2(-1)^s \int_0^\infty \frac{y^s}{s!} K(y)\, dy = \int \frac{y^s}{s!} K(y)\, dy \ .$$

The last kind of kernel needed is called a **mollifier**, because of its exceptional smoothing properties. The term is commonly used in mathematical texts, see e.g. Adams (1975). Mollifiers are class 0 kernels, nonnegative, and zero outside $[-1,1]$. They also have infinitely many continuous derivatives. An example of a mollifier is

$$K(x) = C\ e^{-\frac{1}{1-x^2}} \ , \quad |x| \leq 1$$

(see figure 7.4), where $C$ is a normalization constant.

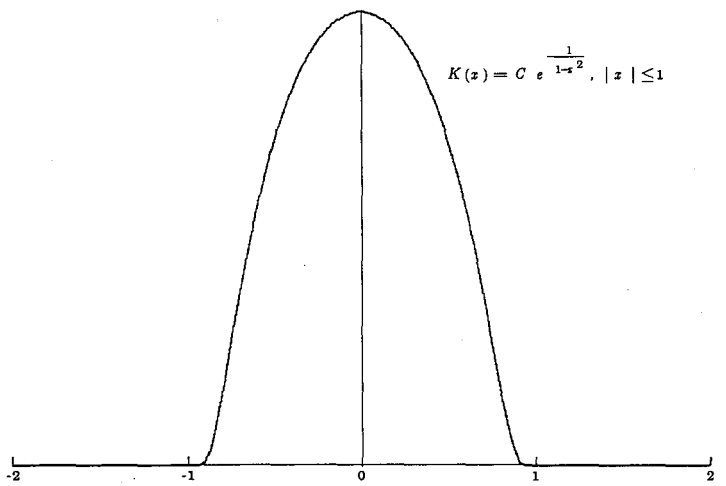


**Figure 7.4.**
A mollifier.

## 7.3. UNIVERSAL DERIVATIVES AND MOLLIFIERS.

Following Butzer and Nessel (1971), we say that $f$ is absolutely continuous when there exists a function $g \in L_1$ such that $f(x) = \int_{-\infty}^{x} g$. In that case, we have $g = f'$ almost everywhere. When a density $f$ has $s-1$ absolutely continuous derivatives, then $f^{(s)}$ exists almost everywhere, so that the functional

$$D_s = \int |f^{(s)}|$$

is well defined for $s = 0,1,2,\ldots$. We run into problems when we study rates of convergence for densities for which the stated condition is not satisfied, unless we can somehow properly generalize $D_s$ without explicitly using the (non-existent) values of $f^{(s)}$. We define

$$D_s^* \stackrel{\Delta}{=} \liminf_{h \downarrow 0} \int |(f * K_h)^{(s)}|$$

where $K$ is a mollifier. Such a definition would make sense if three conditions are satisfied:

A. $D_s^*$ is well-defined for all $f$.

B. Its value is independent of the form of the mollifier $K$.

C. For densities $f$ with $s-1$ absolutely continuous derivatives ($\int |f^{(s)}| < \infty$), $D_s^* = D_s$, i.e. the quantity coincides with the usual functional of $f^{(s)}$.

In this section, we would merely like to point out that conditions A and C are satisfied. It is a bit harder to show that B is also valid. This will be done in the next section for the value $s = 2$ only. One of the interesting properties of convolutions is that $K \in C^\infty$ ($K$ is a mollifier) implies that

$$K_h * \phi \in C^\infty$$

for all functions $\phi \in L_1$. Furthermore, by Young's inequality,

$$\int |K_h * \phi| \leq \int |\phi| < \infty .$$

Finally, $(K_h * \phi)^{(s)} = (K_h)^{(s)} * \phi$. Thus, $D_s^*$ is indeed well-defined. To show that condition C holds, consider $f$ with $s-1$ absolutely continuous derivatives. Then $(K_h * f)^{(s)} = K_h * f^{(s)}$, and thus

$$\int |K_h * f^{(s)}| \leq \int |f^{(s)}|$$

and

$$\int |K_h * f^{(s)}| \to \int |f^{(s)}|$$

as $h \downarrow 0$, a property shared by all kernels when $f^{(s)} \in L_1$.

## 7.4. THE BIAS OF THE KERNEL ESTIMATE FOR CLASS $s$ KERNELS.

We will use Taylor's expansion in this section to derive relationships between $D_s^*$ and the bias term of the kernel estimate when a class $s$ kernel is used. This will be done in several steps. We begin first with upper bounds for the bias:

**Theorem 7.1. Upper bounds for the bias.**

Assume that $K$ is a kernel of order at least $s$, and that $L$ is the kernel with parameter $s$ associated with $K$. For all densities $f$,

$$\int | f*K_h - f | \leq h^s \int | L | D_s^*(f).$$

If $f$ has $s-1$ absolutely continuous derivatives, then

$$\int | f*K_h - f | \leq h^s \int | L | D_s(f).$$

Let $K$ have order greater than $s$. If $f \in W(s,\alpha,C)$ for some $s \geq 1$ (i.e. $f$ vanishes off [0,1], has $s-1$ absolutely continuous derivatives and $f^{(s)}$ is Lipschitz $(\alpha,C)$), then

$$\int | f*K_h - f | \leq 2Ch^{s+\alpha} \int | x |^\alpha | L(x) | \, dx .$$

If $f \in W(0,\alpha,C)$ (i.e. $f$ vanishes off [0,1], and $f$ is Lipschitz $(\alpha,C)$), then

$$\int | f*K_h - f | \leq 2Ch^\alpha \int | x |^\alpha | K(x) | \, dx .$$

**Proof of Theorem 7.1.**

Let us start with the case that $f$ has $s-1$ absolutely continuous derivatives. Then, by Taylor's series expansion with remainder,

$$f(x+y) - f(x) = \sum_{j=1}^{s-1} \frac{y^j}{j!} f^{(j)}(x) + \int_x^{x+y} \frac{(x+y-u)^{s-1}}{(s-1)!} f^{(s)}(u) \, du$$

so that, for class $s$ kernels $K$,

$$\begin{aligned} f*K_h - f &= \int (f(x+y) - f(x)) K_h(y) \, dy \quad (\text{recall that } \int K = 1) \\ &= \sum_{j=1}^{s-1} 0 + \int \int_x^{x+y} \frac{(x+y-u)^{s-1}}{(s-1)!} f^{(s)}(u) \, du \; K_h(y) \, dy \\ &= \int_x^\infty f^{(s)}(u) \int_{u-x}^\infty \frac{(x+y-u)^{s-1}}{(s-1)!} K_h(y) \, dy \, du \\ &\quad - \int_{-\infty}^x f^{(s)}(u) \int_{-\infty}^{u-x} \frac{(x+y-u)^{s-1}}{(s-1)!} K_h(y) \, dy \, du \end{aligned}$$

$$= \int_x^\infty f^{(s)}(u)\,(-1)^s\,(L)_h(u-x)\,du - \int_{-\infty}^x f^{(s)}(u)\,(-1)(-1)^s\,(-1)^s\,(L)_h(x-u)\,du$$

$$= \int_{-\infty}^\infty f^{(s)}(u)\,(L)_h(x-u)\,du = f^{(s)} * (L)_h = h^s f^{(s)} * L_h ,$$

where $(L)_h$ is the kernel associated with $K_h$ and $L$ is the kernel associated with $K$. Therefore, by Young's inequality,

$$\int | f*K_h - f | \le h^s \int | f^{(s)} * L_h | \le h^s \int | L | \int | f^{(s)} | .$$

For general $f$, we first pick a mollifier $\phi$ (we can't use the notation $K$, since $K$ is reserved for our class $s$ kernel), and note that by the Lebesgue density theorem,

$$| f*K_h - f | = \liminf_{a \downarrow 0} | (f*K_h - f)*\phi_a |$$

for almost all $x$. Thus,

$$\int | f*K_h - f | \le \liminf_{a \downarrow 0} \int | (f*K_h - f)*\phi_a | \quad \text{(Fatou's lemma)}$$

$$= \liminf_{a \downarrow 0} \int | (f*\phi_a)*K_h - (f*\phi_a) |$$

$$\le h^s \int | L | \liminf_{a \downarrow 0} \int | (f*\phi_a)^{(s)} | \quad (f*\phi_a \in C^\infty)$$

$$= h^s \int | L | D_s{}^*(f) \quad \text{(definition)} .$$

For the remainder of the proof we assume that the order of $K$ is greater than $s$ (hence, $\int L = 0$ for the parameter $s$ associated kernel $L$). Assume also that $s \ge 1$. If $f$ is in $W(s,\alpha,C)$, then

$$f(x+y) - f(x) = \sum_{j=1}^{s-1} \frac{y^j}{j!} f^{(j)}(x) + \int_x^{x+y} \frac{(x+y-u)^{s-1}}{(s-1)!} \left( f^{(s)}(u) - f^{(s)}(x) \right) du$$

$$- \int_x^{x+y} \frac{(x+y-u)^{s-1}}{(s-1)!}\, du\; f^{(s)}(x) .$$

Thus,

$$f*K_h - f = h^s f^{(s)}(x)\, 1*L_h + h^s (f^{(s)}(\cdot) - f^{(s)}(x))*L_h ,$$

and using the fact that $\int L = 0$, and thus $1*L_h = 0$, we have

$$\int | f*K_h - f | \le h^s \int | \int (f^{(s)}(y) - f^{(s)}(x)) L_h(x-y)\, dy |\, dx .$$

The inner integral has a zero integrand unless $0 \le y \le 1$ or $0 \le x \le 1$. Let us define sets $I_1 = [0,1] \times [0,1]$, $I_2 = [0,1] \times [0,1]^c$, and $I_3 = [0,1]^c \times [0,1]$. By our assumptions on $f^{(s)}$,

$$\int | f*K_h - f | \le h^s \left( \int_{I_1} + \int_{I_2} + \int_{I_3} \right) C\, | x-y |^\alpha\, | L_h(x-y) |\, dx\, dy .$$

Now, the integrals over $I_1$ and $I_2$ taken together do not exceed

$$\int_0^1\int Ch^\alpha \mid x-y \mid^\alpha \mid L(x-y) \mid \; dy \; dx \leq Ch^\alpha \int \mid x \mid^\alpha \mid L(x) \mid \; dx \; .$$

The integrals over $I_1$ and $I_3$ taken together are bounded by the same expression, by symmetry. Therefore,

$$\int \mid f * K_h - f \mid \leq 2Ch^{s+\alpha} \int \mid x \mid^\alpha \mid L(x) \mid \; dx \; .$$

Let us finally handle the case $s=0$. Then, $\mid f(x+y)-f(x) \mid \leq C \mid y \mid^\alpha$, and by an argument similar to the one used for $s>0$,

$$\begin{aligned} &\int \mid \int f(x+y) K_h(y) \, dy - f(x) \mid \; dx \\ &\leq \int\int_{0\leq x+y\leq 1, \text{or } 0\leq x\leq 1} \mid f(x+y)-f(x) \mid \mid K_h(y) \mid \; dy \; dx \\ &\leq \int\int_{0\leq x+y\leq 1, \text{or } 0\leq x\leq 1} C \mid y \mid^\alpha \mid K_h(y) \mid \; dy \; dx \\ &\leq \int_0^1\int C \mid y \mid^\alpha \mid K_h(y) \mid \; dy \; dx + \int_0^1\int C \mid z-x \mid^\alpha \mid K_h(z-x) \mid \; dx \; dz \\ &= 2Ch^\alpha \int \mid y \mid^\alpha \mid K(y) \mid \; dy \; . \blacksquare \end{aligned}$$

The first two results of Theorem 7.1 apply to class $s$ kernels. They basically imply that for finite $D_s^*(f)$, the bias is $O(h^s)$. One might wonder whether for smoother densities $f$, the bias does not tend to zero faster with $h$. If the order of $K$ is $s$, and thus $\int x^s K(x) \neq 0$, then this rate is optimal: no matter how smooth $f$ is, we are in fact "stuck" with the rate $h^s$ for the bias. In other words, the kernel itself limits the performance. At first glance therefore, it seems important to insure that the order of $K$ be as large as possible. In fact, superkernels have infinite order, and show the capability of adapting to nearly any degree of smoothness of $f$. Let us give some results about the built-in limitations of kernels of finite order $s$:

**Theorem 7.2.**

Let $K$ be a kernel of order $s$. If $f$ has $s-1$ absolutely continuous derivatives, then

$$\lim_{h \downarrow 0} \frac{\int | f * K_h - f |}{h^s \left| \int \frac{x^s}{s!} K(x)\, dx \right|} = D_s(f).$$

Observe that the integral in the denominator is $\int L$ where $L$ is the associated kernel with parameter $s$.

The statements following now are valid for all densities $f$. For a class 2 kernel with $K \geq 0$,

$$\lim_{h \downarrow 0} \frac{\int | f * K_h - f |}{h^2 \int \frac{x^2}{2} K(x)\, dx} = D_s^*(f).$$

In other words, the definition of $D_2^*(f)$ is independent of the choice of a mollifier. Furthermore, for kernels $K$ of order $s$ possessing compact support, and densities $f$ with $D_{s+1}^*(f) < \infty$,

$$\lim_{h \downarrow 0} \frac{\int | f * K_h - f |}{h^s \left| \int \frac{x^s}{s!} K(x)\, dx \right|} = D_s^*(f).$$

If $D_{s+1}^*(f) = \infty$, and $K$ has order $s$ (but possibly infinite support), we still have

$$\liminf_{h \downarrow 0} \frac{\int | f * K_h - f |}{h^s \left| \int \frac{x^s}{s!} K(x)\, dx \right|} \geq D_s^*(f).$$

**Proof of Theorem 7.2.**

Assume first that $f$ has $s-1$ absolutely continuous derivatives. Since $\int | f^{(s)} | < \infty$ by assumption, we have

$$\frac{\int | f * K_h - f |}{h^s} = \int | f^{(s)} * L_h |$$

$$= \left| f^{(s)} \int L + \int (f^{(s)} * L_h - f^{(s)} \int L) \right|.$$

Since $\int | L | < \infty$, we have

$$\int | f^{(s)} * L_h - f^{(s)} \int L | \to 0.$$

Therefore,

$$\frac{\int | f *K_h - f |}{h^s} \to \int | f^{(s)} | \; | \int L | \; .$$

Let us now look at arbitrary $f$ with finite $D_s{}^*(f)$ (the case of infinite values can be handled by a large compact set argument). Then,

$$\int | f *K_h - f | / h^s \geq \int | (f * \phi_a) * K_h - (f * \phi_a) | / h^s$$

($\phi$ is a mollifier, $a$ is arbitrary)

$$\to \int | (f * \phi_a)^{(s)} | \; | \int L | \quad (\text{as } h \downarrow 0 \text{, by the first part}) \; .$$

Thus,

$$\liminf_{h \downarrow 0} \int | f *K_h - f | / h^s \geq \sup_{a > 0} \int | (f * \phi_a)^{(s)} | \; | \int L | \geq D_s{}^*(f) \; | \int L | \; .$$

The upper bound of Theorem 7.1, and the lower bound proved here agree when $K \geq 0$ and is in class 2, since $| \int L | = \int | L | = \int (x^2/2) K(x) \, dx$. There are other instances in which they agree. Assume for example that $D_{s+1}{}^*(f) < \infty$ and that $K$ has compact support. Then

$$\int | f *K_h - f | / h^s$$

$$\leq \int | f - f * \phi_a | / h^s + \int | f * \phi_a - f * \phi_a *K_h | / h^s + \int | f * \phi_a *K_h - f *K_h | / h^s$$

$$\leq 2 \int | f - f * \phi_a | / h^s + \int | (f * \phi_a)^{(s)} * L_h | \; .$$

The last term is not greater than

$$\int | (f * \phi_a)^{(s)} | \; | \int L | + \int | (f * \phi_a)^{(s)} * L_h - (f * \phi_a)^{(s)} \int L |$$

$$\leq \int | (f * \phi_a)^{(s)} | \; | \int L | + \int | (f * \phi_a)^{(s+1)} | \; h \int | L | \int | M |$$

where $M$ is the kernel of parameter $s+1$ associated with $L / \int L$. If we let $a \downarrow 0$, then the upper bound for $\int | f *K_h - f | / h^s$ tends to

$$0 + D_s{}^*(f) \; | \int L | + D_{s+1}{}^*(f) h \int | L | \; .$$

provided that $\int | M | < \infty$. This is certainly the case when $K$ has compact support. Next, let $h \downarrow 0$, so that this upper bound agrees with the lower bound established above. ■

## 7.5. SATURATION AND UNBIASEDNESS.

It will help to consider the following simple table

| KERNEL : | ORDER 2 | ORDER 4 | ORDER $s$ | ORDER $\infty$ |
|---|---|---|---|---|
| SMOOTHNESS OF $f$ | | | | |
| No restriction | $o(1)$ | $o(1)$ | $o(1)$ | $o(1)$ |
| $D_2^*(f)<\infty$ | $\approx h^2$ | $o(h^2)$ | $o(h^2)$ | $o(h^2)$ |
| $D_4^*(f)<\infty$ | $\approx h^2$ | $\approx h^4$ | $o(h^4)$ | $o(h^4)$ |
| $D_s^*(f)<\infty$ | $\approx h^2$ | $\approx h^4$ | $\approx h^s$ | $o(h^s)$ |
| Bounded spectrum | $\approx h^2$ | $\approx h^4$ | $\approx h^s$ | $o(h^s)$ for any $s$ |

$$\text{Bias} \int | f*K_h - f | .$$

It should be noted here that the bias is $o(1)$ in all cases, a fact discussed at length in chapter 2. For order 2 kernels, the bias can decrease at the rate $h^2$ but not faster; there is not a single exception. In particular, this rate is best possible for all nonnegative kernels. This phenomenon is known as **saturation** (for more on saturation, consult Butzer and Nessel, 1971). A similar thing happens for order $s$ kernels, only at the rate $h^s$. Order $s$ kernels do not allow the bias to tend to zero faster than $h^s$. Interestingly, superkernels adapt automatically to the smoothness of $f$ as long as smoothness is measured in terms of the polynomial rate of convergence of the bias (in $h$).

The smoothest densities are perhaps those with bounded spectrum, i.e. the densities whose characteristic function vanishes outside a compact set $[-T, T]$. Let us call this class $BS_T$. Densities in this class are necessarily analytic (since the characteristic function is $O(e^{-c|t|})$ as $|t| \to \infty$; see e.g. Kawata (1972, p. 439)). For these densities, it is possible to have zero bias, as we will now see. Assume that $K$ is a flattop superkernel (for the existence of such kernels, see exercise 7.2) with characteristic function $\psi$ where $\psi = 1$ on $[-1,1]$ (without loss of generality). If $f$ has characteristic function $\phi$, then $f*K_h - f$ has characteristic function

$$\phi(t)\Big(\psi(th) - 1\Big) .$$

This is identically zero (and hence $\int | f*K_h - f | = 0$) if $\psi(th) = 1$ for all $|t| \le T$. This is in turn satisfied if $h \le 1/T$. We have thus shown the remarkable fact that

$$f * K_h \equiv f$$

for $f \in BS_T$, $h \leq \frac{1}{T}$, and any flattop kernel $K$ with characteristic function of value 1 on [-1,1]. Thus, for flattop kernels, we can truly state that they adapt to the smoothness of $f$ for all $f$. Superkernels without the flattop property do not necessarily yield unbiasedness on $BS_T$. See exercise 7.3.

## 7.6. THE VARIATION OF THE KERNEL ESTIMATE.

In this section, we relate the variation $E(\int | f_n - f * K_h |)$ to two quantities, $\int \sqrt{f}$, a measure of the heaviness of the tail of $f$, and $\int K^2$, a measure of how large $K$ is. We have seen that the bias term is small when $f$ is smooth and $K$ has a large order $s$ (i.e. $K$ oscillates a lot). Unfortunately, smooth densities smear their mass out in such a way that $\int \sqrt{f}$ is large, and oscillating kernels have generally speaking larger values $\int K^2$ than monotone kernels with similar widths. Thus, we already have a preview of the compromise ahead of us: what is good for the bias is bad for the variation, and vice versa.

**Theorem 7.3.**

Let $f_n$ be the kernel estimate with kernel $K$.

$$E(\int | f_n - f * K_h |) \leq \int \sqrt{Var(f_n - f * K_h)} \leq \frac{\int \sqrt{f * (K^2)_h}}{\sqrt{nh}} .$$

**Proof of Theorem 7.3.**

Use the Cauchy-Schwarz inequality and the following inequality:

$$E((f_n - f * K_h)^2) = \frac{1}{n^2} n E((K_h(x - X_1) - f * K_h)^2)$$

$$\leq \frac{1}{n} E(K_h{}^2(x - X_1))$$

$$= \frac{1}{nh} (K^2)_h * f(x) . \blacksquare$$

For small $h$, the upper bound of Theorem 7.3 is roughly speaking equal to a constant times $\int\sqrt{f}\,/\sqrt{nh}$. The $1/\sqrt{nh}$ term is due to the effect of the central limit theorem because locally, $f_n$ is an average of about $nh$ observations. The **tail factor** $\int\sqrt{f}$ is somehow related to $E\,|X|$ where $|X|$ has density $f$. In most cases, the tail factor and $E\,|X|$ are finite or infinite together. See however exercise 7.5.

To relate $\int\sqrt{f}$ more precisely to the tail of $f$, we need

**Lemma 7.1. Carlson's inequality (Carlson, 1934).**

Let $g \geq 0$ be a measurable function on the real line. Then

$$\int_0^\infty g \leq \sqrt{\pi}\left(\int_0^\infty g^2\right)^{\frac{1}{4}}\left(\int_0^\infty x^2g^2\right)^{\frac{1}{4}},$$

$$\int_{-\infty}^\infty g \leq \sqrt{2\pi}\left(\int_{-\infty}^\infty g^2\right)^{\frac{1}{4}}\left(\int_{-\infty}^\infty x^2g^2\right)^{\frac{1}{4}}.$$

**Proof of Lemma 7.1.**

We will use the Cauchy-Schwarz inequality, and the fact that $\int_{-\infty}^{\infty}(1+x^2)^{-1}=\pi$:

$$\begin{aligned}\int g &= \int\frac{\sqrt{g^2+a^2x^2g^2}}{\sqrt{1+a^2x^2}} \quad (a\in R)\\ &\leq \sqrt{\int(g^2+a^2x^2g^2)\int(1+a^2x^2)^{-1}}\\ &= \sqrt{\frac{\pi}{2a}\left(\int g^2+a^2\int x^2g^2\right)}\\ &= \sqrt{\pi\sqrt{\int g^2}\sqrt{\int x^2g^2}}\end{aligned}$$

If we choose the optimal value for $a$, i.e.

$$a = \sqrt{\frac{\int g^2}{\int x^2g^2}}.$$

All integrals were from 0 to $\infty$. ■

If we take $g = \sqrt{f}$ in Carlson's inequality, then we have

$$\int \sqrt{f} \leq \sqrt{2\pi} E^{\frac{1}{4}}(X^2)$$

where $X$ has density $f$. Since the left-hand side is translation-invariant, we obtain

**Lemma 7.2.**

For any density $f$ with variance $\sigma^2$,

$$\int \sqrt{f} \leq \sqrt{2\pi\sigma}.$$

When $f$ is regularly varying of order $r$, i.e.

$$\lim_{x \to \infty} \frac{f(tx)}{f(x)} = t^r$$

for all $t > 0$, and $f = 0$ on the negative halfline, we have $\int \sqrt{f} = \infty$ if $r > -2$ and $\int \sqrt{f} < \infty$ if $r < -2$. This again relates the finiteness of the tail factor to that of $\int |x| f$. The finiteness of $\int \sqrt{f}$ is important for us because of Theorem 7.3 and

**Lemma 7.3.**

If $K$ is a square integrable kernel, then

$$\int \sqrt{f * (K^2)_h} \geq \int \sqrt{f} \sqrt{\int K^2}.$$

**Proof of Lemma 7.3.**

$$\sqrt{f * (K^2)_h} / \sqrt{\int (K^2)_h} \geq \sqrt{f} * (K^2)_h / \int K^2$$

by Jensen's inequality. ■

As a corollary, we see that $\int \sqrt{f} = \infty$ makes the bound of Theorem 7.3 useless. Of course, for this to happen, $f$ needs to have a fat tail; Carlson's

inequality tells us that at the very least, the variance must be infinite. Unfortunately, as shown in exercise 7.6, the upper bound of Theorem 7.3 can be infinite ($\int f*(K^2)_h = \infty$ for all $h>0$) even if $\int\sqrt{f} < \infty$. Furthermore, the bound obtained in Theorem 7.3 is not loose, because the $1/\sqrt{nh}$ rate of convergence for the variation is impossible to achieve when $\int\sqrt{f} = \infty$ (see exercise 7.7). It is therefore necessary to analyze the upper bound of Theorem 7.3 a bit further. We will do this by establishing yet another explicit bound, since explicit bounds are needed to prove the minimax optimality of the kernel estimate later on.

**Theorem 7.4.**

Let $f$ be a density with variance $\sigma^2<\infty$. Assume that $K$ is a kernel for which $\int(1+x^2)K^2<\infty$. Then

$$E\left(\int |f_n - f*K_h|\right) \le (1+o(1))\frac{\sqrt{\int K^2}\int\sqrt{f}}{\sqrt{nh}}$$

if $h\to 0$ as $n\to\infty$. The $o(1)$ term can be taken equal to

$$\left(2\sigma^2+h^2\frac{\int x^2K^2}{\int K^2}\right)^{\frac{1}{4}}\frac{\sqrt{2\pi}}{\int\sqrt{f}}\int^{\frac{1}{4}}|f*Q_h - f|,$$

where $Q = K^2/\int K^2$ is the **quadratic associated kernel.**

**Proof of Theorem 7.4.**

That the $o(1)$ term defined in the statement of the theorem is indeed $o(1)$ follows from the fact that all kernels are approximate identities for all densities. We will also need the fact that $\int\sqrt{f} < \infty$ (which follows from the finiteness of $\sigma$, and Lemma 7.2). We begin with the upper bound for the variation given in Theorem 7.3:

$$\frac{\sqrt{\int K^2}\int\sqrt{f*Q_h}}{\sqrt{nh}}.$$

Note that

$$\int\sqrt{f*Q_h} \le \int\sqrt{f} + \int\sqrt{|f-f*Q_h|}$$

$$\le \int\sqrt{f} + \sqrt{2\pi}\left(\int|f-f*Q_h|\int x^2|f-f*Q_h|\right)^{\frac{1}{4}} \quad \text{(Carlson's inequality)}$$

$$\le \int\sqrt{f} + \sqrt{2\pi}\left(\int|f-f*Q_h|(\sigma^2+\sigma^2+h^2\textstyle\int x^2K^2/\int K^2)\right)^{\frac{1}{4}}. \quad \blacksquare$$

## 7.7. MINIMAX UPPER BOUNDS.

The explicit bounds for the bias and variation can now be put to good use. They can be combined to yield upper bounds for the expected $L_1$ error, uniformly over some class of densities. We will consider the class $\mathbf{F}=W(s,\alpha,C)$ only. We inherit the following notation from the previous sections: $L$ is the parameter $s$ associated kernel, $K$ is assumed to have order greater than $s$, $\sigma^2=\int x^2 f$ ($\leq 1$). In addition, we will use the fact that $\int\sqrt{f} \leq \sqrt{\int f} =1$, and we will assume that

$$\mu_\alpha = \int |x|^\alpha |L(x)|\, dx < \infty .$$

Using

$$E\left(\int |f-f_n|\right) \leq \int |f-f*K_h| + E\left(\int |f*K_h - f_n|\right),$$

we obtain the following inequality from Theorems 7.1 and 7.4:

$$E\left(\int |f-f_n|\right) \leq 2Ch^{s+\alpha}\mu_\alpha + \frac{\sqrt{\int K^2}\int\sqrt{f}}{\sqrt{nh}}$$

$$+ \frac{\sqrt{2\pi\int K^2}}{\sqrt{nh}}\left(2\sigma^2 + h^2\frac{\int x^2K^2}{\int K^2}\right)^{\frac{1}{4}}\left(\sup_{\mathbf{F}}\int |f*Q_h - f|\right)^{\frac{1}{4}}$$

which can be put in the form

$$(1+o(1))\left(Ah^{s+\alpha} + \frac{B}{\sqrt{nh}}\right)$$

if we choose $h$ such that $h\to 0$ and $nh\to\infty$. For the verification of the fact that $\sup_{\mathbf{F}}\int |f*Q_h - f| = o(1)$, see exercise 7.13. The constants $A$ and $B$ can be taken as follows:

$$A \stackrel{\Delta}{=} 2C\mu_\alpha\,,\quad B \stackrel{\Delta}{=} \sqrt{\int K^2}\,.$$

The functions $Ah^{s+\alpha}$, $B/\sqrt{nh}$ and their sum are shown in figure 7.5.

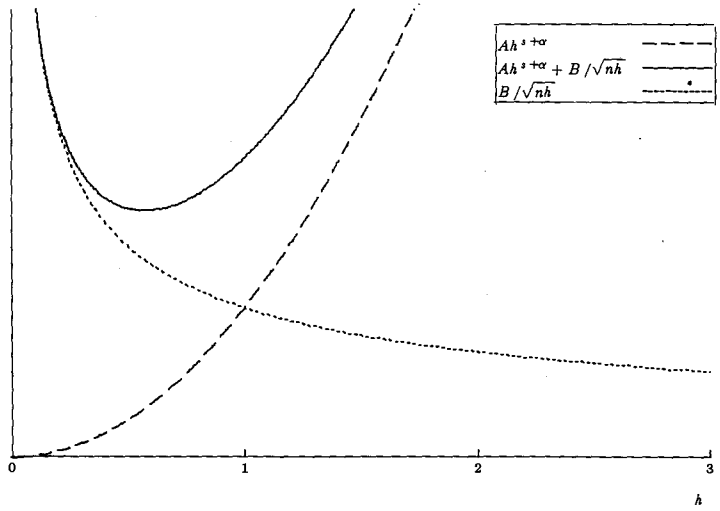


**Figure 7.5.**
Function to be minimized with respect to $h$.

The sum of the functions shown in figure 7.5 is minimal when

$$A\,(s+\alpha)h^{s+\alpha+1} - \frac{B}{2\sqrt{n}\,h^{3/2}} = 0\,,$$

i.e. when

$$h = \left(\frac{B}{2A\,(s+\alpha)\sqrt{n}}\right)^{\frac{1}{s+\alpha+\frac{1}{2}}},$$

and the minimal value is

$$C^* \left(AB^{2(s+\alpha)}\right)^{\frac{1}{2(s+\alpha)+1}} n^{-\frac{s+\alpha}{2(s+\alpha)+1}},$$

where

$$C^* = \left(2(s+\alpha)\right)^{\frac{1}{2(s+\alpha)+1}} \left(1 + \frac{1}{2(s+\alpha)}\right).$$

In other words, we have just shown

**Theorem 7.5.**

Let $f_n$ be the kernel estimate with kernel $K$ of order greater than $s$, and with smoothing factor

$$h = \left( \frac{\sqrt{\int K^2}}{4C\,\mu_\alpha\,(s+\alpha)\sqrt{n}} \right)^{\frac{1}{s+\alpha+\frac{1}{2}}} .$$

Then

$$\sup_{W(s,\alpha,C)} E\left(\int | f - f_n |\right) \leq (1+o(1))\; C^*\; n^{-\frac{s+\alpha}{2(s+\alpha)+1}} \left(2C\,\mu_\alpha (\int K^2)^{s+\alpha}\right)^{\frac{1}{2(s+\alpha)+1}}$$

where $C^*$ and $\mu_\alpha$ are defined above.

It is perhaps worth noting that the minimax lower bound for the Lipschitz classes, which is of the form

$$(C_1(s,\alpha)+o(1)) \left( \frac{C}{n^{s+\alpha}} \right)^{\frac{1}{2(s+\alpha)+1}} ,$$

depends in the same manner upon $C$ and $n$ as does the upper bound for the kernel estimate given here. It should also be stressed that the kernel estimate is completely defined in terms of $n, s, \alpha, C$. For $W(s,\alpha,C)$, the given kernel estimate is minimax-optimal. Other more sophisticated estimates can at best improve the constant in the upper bound. This brings us to the issue of the best form of $K$, i.e. the form that minimizes the upper bound.

## 7.8. THE OPTIMAL KERNEL.

Let us try to find out for which kernel(s) the minimax upper bounds of the previous section are minimal. What we need to minimize is

$$\left(\int K^2\right)^{s+\alpha} \int | x |^\alpha | L(x) |\; dx .$$

Here $K$ is a kernel of order greater than $s$, and $L$ is the parameter $s$ kernel associated with $K$. We will consider only two cases, $s=0$ and $s=1$. For $s=0$, $L=K$, so that we need to minimize

$$\left(\int K^2\right)^\alpha \int | x |^\alpha | K | .$$

We have

**Lemma 7.4.**

For all kernels $K$ of order at least equal to one,

$$\left(\int K^2\right)^\alpha \int |x|^\alpha |K| \geq \frac{1}{2\alpha+1}\left(\frac{\alpha+1}{2\alpha+1}\right)^\alpha .$$

Equality is reached for

$$K_0(x) = \frac{\alpha+1}{2\alpha}(1-|x|^\alpha)_+ .$$

**Proof of Lemma 7.4.**

We first observe that

$$\int K_0 = 1 , \int |x|^\alpha K_0 = \frac{1}{2\alpha+1} , \int K_0^2 = \frac{\alpha+1}{2\alpha+1} ,$$

and that $K_0$ is symmetric. The product in the statement of the lemma is scale invariant, so we need only consider kernels for which $\int |x|^\alpha K \equiv 1/(2\alpha+1)$. For any other kernel of order at least one, $\int (K-K_0)=0$, and $\int |x|^\alpha (K-K_0)=0$. Therefore,

$$\int K^2 = \int (K-K_0)^2 + \int K_0^2 + 2\int_{-1}^{1} \frac{\alpha+1}{2\alpha}(1-|x|^\alpha)(K-K_0)$$

$$= \int (K-K_0)^2 + \int K_0^2 + 2\int_{[-1,1]^c} \frac{\alpha+1}{2\alpha}(|x|^\alpha - 1)(K-K_0)$$

$$\geq \int (K-K_0)^2 + \int K_0^2$$

when $K$ is a density (the integrand of the last term in the middle expression is nonnegative on $[-1,1]^c$). Thus, for densities $K$, we have proved the result. For kernels that can be split into a positive and negative part, $K = K_+ - K_-$, we see that

$$\left(\int K^2\right)^\alpha \int |x|^\alpha |K| \geq \left(\int K_+^2\right)^\alpha \int |x|^\alpha |K_+|$$

and this is at least equal to the value of the product at $K_0$ (since $\int K_+ \geq 1$). This concludes the proof of Lemma 7.4. ■

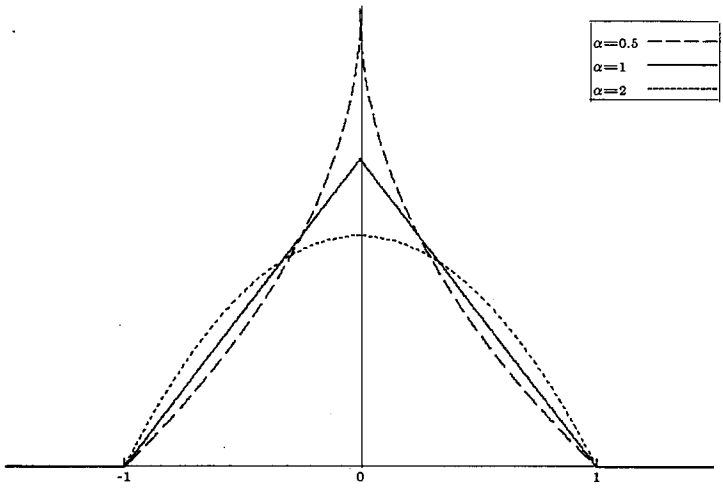


**Figure 7.6.**
Several optimal kernels are shown.

For the ordinary Lipschitz class $W(0,1,C)$, a quick computation shows that with the optimal kernel $(1-x)_+$ (recall that $s+\alpha=1$), and

$$h = \left(\frac{3}{8nC^2}\right)^{\frac{1}{3}},$$

the kernel estimate has the following uniform upper bound:

$$\sup_{f \in W(0,1,C)} E\left(\int |f_n - f|\right) \leq (1+o(1)) \left(\frac{3C}{n}\right)^{\frac{1}{3}}.$$

This should be compared with the lower bound we obtained for this class,

$$(1+o(1))\sqrt{1-\frac{4}{C}} \left(\frac{162\ C}{10000\ n}\right)^{\frac{1}{3}}.$$

As $n \to \infty$, the ratio of upper to lower bound tends to

$$\left(\frac{30000}{1296}\right)^{\frac{1}{3}} \approx 2.8499599/\sqrt{1-\frac{4}{C}}.$$

The optimization of Lemma 7.4 is generally applicable when $K \geq 0$. We observe for example that when $L$ is the parameter $s$ kernel associated with a kernel $K \geq 0$, then

$$\begin{aligned}\int |x|^{\alpha} |L(x)| \, dx &= 2\int_0^{\infty} x^{\alpha} \int_x^{\infty} \frac{(y-x)^{s-1}}{(s-1)!} K(y) \, dy \, dx \\ &= 2\int_0^{\infty} K(y) \int_0^{y} \frac{x^{\alpha}(y-x)^{s-1}}{(s-1)!} \, dx \, dy \\ &= 2\int_0^{\infty} K(y) y^{s+\alpha} \int_0^{1} \frac{v^{\alpha}(1-v)^{s-1}}{(s-1)!} \, dv \, dy \\ &= \frac{\Gamma(\alpha+1)}{\Gamma(\alpha+s+1)} \int K(y) |y|^{s+\alpha} \, dy \, .\end{aligned}$$

Since $K$ must have order greater than $s$ for the uniform bounds to be applicable, the computation of $\int |x|^{\alpha} |L|$ is only useful for $s=0$ or $s=1$. We conclude that the optimal kernel for $W(1,\alpha,C)$ has the form

$$\frac{\alpha+s+1}{2(\alpha+s)} (1-|x|^{1+\alpha})_+ \, .$$

For $W(1,1,C)$, the uniform bound obtained with the kernel estimate is asymptotic to a universal constant times $C^{1/5}/n^{2/5}$.

## 7.9. INDIVIDUAL UPPER BOUNDS.

There is a discrepancy between best possible rates of convergence for individual $f$, and uniform rates over classes $\mathbf{F}$. In fact, we can regard individual rates as uniform rates over $\mathbf{F}=\{f\}$. For example, for the kernel estimate, we have very often

$$\frac{\inf_{h>0,K} E\left(\int |f_n - f|\right)}{\inf_{h>0,K} \sup_{f \in \mathbf{F}} E\left(\int |f_n - f|\right)} \to 0 \quad \text{as } n \to \infty \, ,$$

where $\mathbf{F}$ is a class containing $f$. In other words, for most densities, we can do much better than what we might think is possible merely by studying minimax results. In one case however, there is relatively little difference between minimax upper bounds and individual upper bounds, i.e. when $K$ is nonnegative.

**Theorem 7.6.**

Assume that $f$ is a density with $D_2^*(f)<\infty$, and $\int x^2 f<\infty$. If $K$ is a nonnegative order 2 kernel, for which $\int(1+x^2)K^2<\infty$, then

$$\inf_{h>0} E\left(\int |f_n - f|\right) \leq (1+o(1))\,(2^{\frac{1}{5}}+2^{-\frac{9}{5}})\,\phi(K)\,\psi(f)\,n^{-\frac{2}{5}},$$

where

$$\phi(K) \stackrel{\Delta}{=} \left(\int x^2 K \left(\int K^2\right)^2\right)^{\frac{1}{5}},$$

and

$$\psi(f) \stackrel{\Delta}{=} \left(D_2^*(f)\int \sqrt[4]{f}\right)^{\frac{1}{5}}.$$

The inequality is valid for example when

$$h = n^{-\frac{1}{5}} \frac{\left(\int K^2\right)^{\frac{1}{5}}\left(\int \sqrt{f}\right)^{\frac{2}{5}}}{\left(2D_2^*(f)\int x^2 K\right)^{\frac{2}{5}}}.$$

**Proof of Theorem 7.6.**

If $h \to 0$ and $nh \to \infty$, we have

$$E\left(\int |f_n - f|\right) \leq (1+o(1))h^2 D_2^*(f)\frac{1}{2}\int x^2 K + (1+o(1))\frac{\sqrt{\int K^2}\int\sqrt{f}}{\sqrt{nh}}.$$

Here we used Theorems 7.1 and 7.4. If we ignore the $o(1)$ terms, then we are left with a simple minimization problem with respect to $h$. The "best" $h$ is given in the statement of the Theorem, and satisfies the conditions $h \to 0$ and $nh \to \infty$. The corresponding upper bound is the one given in the statement of the theorem. ■

The upper bound deserves some special attention. First of all, it depends upon a product of three factors, $n^{-2/5}$, $\phi(K)$ and $\psi(f)$. The dependence upon $n$ is the same one achieved in the minimax theory for the (Bretagnolle-Huber) class of all densities on [0,1] for which $f$ and $f'$ are absolutely continuous, and $\psi(f)\leq r$ for some constant $r$. We have already hinted that

$$\inf_{h>0} E\left(\int |f_n - f|\right) \geq (C + o(1))\, \phi(K)\, \psi(f)\, n^{-\frac{2}{5}},$$

for some constant $C$, and **all** densities $f$. Thus, the upper bound is fine modulo a small universal constant. The factor $\psi(K)$ is minimal for **Bartlett's kernel** (Bartlett, 1962)

$$K(x) = \frac{3}{4}(1-x^2)_+,$$

and the minimal value is $\frac{1}{5}(\frac{2}{5})^{2/5}$ (see Lemma 7.4). Finally, the factor $\psi(f)$ could truly be called the **difficulty factor** for kernel estimates, since it seems to indicate how hard it is to estimate $f$ with a (nonnegative) kernel estimate. It is obviously scale and translation-invariant, depending only upon the shape of $f$.

$\psi(f)$ is infinite for one of two reasons: either the tail of $f$ is too big ($\int \sqrt{f} = \infty$), as is the case for the Cauchy density, or $f$ oscillates too much ($D_2^*(f) = \infty$). Examples of densities with $D_2^*(f) = \infty$ include all densities with a simple discontinuity (such as the uniform density on [0,1] or the exponential density), and all densities with an infinite peak (ess sup $f(x) = \infty$). Not all finite values for $\psi(f)$ are possible either. It is known (Devroye and Penrod, 1984) that

$$\inf_f D_2^*(f) = \frac{4}{3^{4/5}} \approx 1.66097458615.$$

The infimum is attained for the isosceles triangular density $(1 - |x|)_+$ and all its shifted and rescaled versions. Consider for example the plane formed by $D_2^*(f)$ as the $x$-axis, and $\int \sqrt{f}$ as the $y$-axis (figure 7.7). The region below the curve $xy^4 \leq c$ ($c$ is some constant) is forbidden. The isosceles triangular density moves on the border of the forbidden region as its scale changes. The lower bound for $\psi(f)$ is really due to the fact that when one has to draw a density, one either needs to create a big tail if the density is to be smooth, or one needs a lot of oscillation if the tail is to be small.

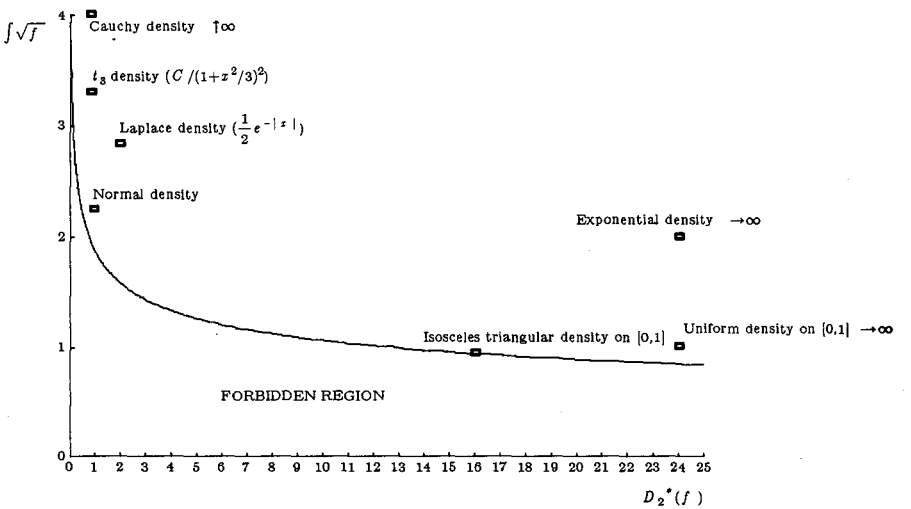


**Figure 7.7.**
Plane of $\int\sqrt{f}$ versus $D_2^*(f)$.

If a given density satisfies the conditions of Theorem 7.6, i.e. the value for $\psi(f)$ is finite, and the variance is finite, then by taking an order 4 kernel, it is relatively easy to establish that the bias is $o(h^2)$, so that

$$\inf_{h>0} E\left(\int |f_n - f|\right) = o\left(n^{-\frac{2}{5}}\right)$$

but it is difficult to give an explicit formula for $h$ in terms of $f$, unless one assumes more (e.g., $D_4^*(f)<\infty$). Essentially, as observed as early as 1962 by Bartlett, and rediscovered later by others, if $D_s^*(f)<\infty$ for even $s$, and we choose an order $s$ kernel $K$, it is possible to insure that

$$E\left(\int |f_n - f|\right) = O\left(n^{-\frac{s}{2s+1}}\right)$$

by taking $h \sim n^{-1/(2s+1)}$. As $s$ grows, we get better rates of convergence, but these rates apply to a shrinking collection of densities. Usually, we don't know anything about $f$, let alone the fact that $D_s^*(f)$ is finite. Furthermore, if one chooses $h \sim n^{-1/(2s+1)}$, but $f$ is not as smooth as originally thought so that the bias decreases e.g. as $\sim h$, then the particular choice of $h$ is actually disastrous, as the bias alone decreases as $n^{-1/(2s+1)}$ (in this artificial example) instead of $n^{-s/(2s+1)}$.

## 7.10. MODIFIED KERNEL ESTIMATES.

The design of a kernel estimate boils down to the choice of $h$ and $K$. Under certain conditions on $f$, it is possible to give formulas of the form

$$h = \alpha(n)\beta(f)$$

for the smoothing factor so that a relatively good upper bound on the expected $L_1$ error is minimized. Here $\alpha(n)$ is a function of $n$, such as $n^{-1/5}$, and $\beta(f)$ is a functional involving $f$ (see Theorem 7.6). Since $f$ is unknown, $\beta(f)$ is unknown, and we have a vicious circle. There are a few ways around this problem.

Some statisticians suggest making a parametric assumption about $f$, and estimating the parameter or parameters from the data. These estimates are then plugged back into the parametric model $g_n$, which in turn is used to compute $\beta(g_n)$, which hopefully is not too far from the unknown $\beta(f)$. This has worked well in some applications, in which exploratory data analysis had already established that the unknown density could almost be fitted by some parametric method. See e.g. Deheuvels (1977).

Others argue, quite correctly, that the formula $\alpha(n)\beta(f)$ for $h$ is based upon non-verifiable conditions on $f$. They suggest going back to general principles like cross-validation and maximum likelihood in an attempt to obtain the best choice of $h$ directly. Ideally, one would hope to find a function $h = h(X_1, \ldots, X_n)$ such that for all $f$,

$$\frac{E(\int |f_n - f|)}{\inf_{h>0} E(\int |f_{nh} - f|)} \to 1$$

as $n \to \infty$, where $f_{nh}$ is the ordinary kernel estimate with deterministic $h$, and $f_n$ is the kernel estimate with data-based $h$. It is still unknown whether such a formula exists. If it does, we would achieve the asymptotically optimal rate of convergence without actually knowing the rate of convergence. Stone (1984,1985) has obtained a result in this spirit for the expected $L_2$ error of the kernel estimate and the histogram estimate. His data-based smoothing factor uses least-squares cross-validation: for further work and references, consult Hall (1983), Burman (1985), Marron (1985) or Scott and Terrell (1986).

A data-based smoothing factor $h$ supposedly adjusts itself automatically to the smoothness and size of the tail of $f$. For extremely smooth densities, the limiting factor is the order of $K$, because of the saturation phenomenon in the bias of the kernel estimate. It is necessary to jack the order of $K$ up to the maximal smoothness expected in $f$. The issue of the choice of the order of $K$ can be circumvented if one picks a superkernel. The cost of doing this is considerable: the fact that $f_n$ can take negative values and that $K$ has infinite support implies that a lot of computational time is required to obtain $f_{n+}/\int f_{n+}$, the normalized density estimate. Furthermore, the constant term in asymptotic expressions for the expected $L_1$ error (see Theorem 7.6) typically grows as the order of the kernel grows, and is large for superkernels.

One should question whether it is appropriate to spend a lot of time and effort on the choice of $h$ and/or $K$, because doing so is admitting that one is willing to accept the asymptotic performance of the kernel estimate, which we now know is a complicated function of $f$. Recalling that in many cases it is a function of $\psi(f)$, it is perhaps wiser and easier to try to transform the data in such a way that $\psi(f)$ becomes smaller, to estimate the density of the transformed data by the kernel method, and to retransform the density estimate. By the invariance of the $L_1$ distance under continuous one-to-one transformations, it suffices to analyze the expected $L_1$ error for the transformed data, with its presumably better value of $\psi(f)$. Ideally, with nonnegative $K$, one would like to transform the data in such a way that $f$ becomes triangular (see Theorem 7.6 and its discussion). The question remains of how one can construct a good data-based transformation with provably excellent properties. See for example Devroye, Machell and Penrod (1983) or Devroye and Gyorfi (1985).

## 7.11. RATES OF CONVERGENCE WITH SUPERKERNELS.

One of the advantages of a flattop superkernel is that it can be used for all densities; provided that we are able to choose $h$ properly, we can adapt to any degree of smoothness. The variation for superkernels is dealt with in Theorem 7.6: it is not different than for any other kernel.

The bias term can be handled differently. Since $K$ has an infinite order, all the results of Theorem 7.1 apply. This has the drawback that densities have to be dealt with according to membership in classes defined in terms of the $s$-th generalized derivative. There is another way of studying the bias, which provides us with a continuum of rates, based upon the behavior of the characteristic function $\phi$ of $f$.

Assume that $f$ has characteristic function $\phi$ and that $K$ has "characteristic function" $\psi$, where $\psi \equiv 1$ on $[-T,T]$, and that both $f$ and $K$ are in $L_2$. The tools for obtaining upper bounds for $\int |f*K_h - f|$ are essentially the Cauchy-Schwarz inequality or Carlson's inequality, and Parseval's identity (also called Bessel's equality):

A. If $f$ has support contained in [0,1], then.

$$\int_0^1 |f*K_h - f| \le \sqrt{\int_0^1 (f*K_h - f)^2} \quad \text{(Cauchy-Schwarz)}.$$

B.

$$\int |f*K_h - f| \le \sqrt{2\pi} \left(\int (f*K_h - f)^2\right)^{\frac{1}{4}} \left(\int x^2 (f*K_h - f)^2\right)^{\frac{1}{4}}$$

(Carlson's inequality).

C.

$$\int (f*K_h - f)^2 = \frac{1}{2\pi}\int (\phi(t)\psi(th) - \phi(t))^2\, dt \quad \text{(Parseval's identity)}.$$

For our flattop kernel $K$, we deduce that

$$\int (f*K_h - f)^2 \le \frac{1}{2\pi} \int_{|t|>T/h} \phi^2(t)\, dt .$$

This can be plugged back into bounds for the bias:

$$\int_0^1 |f*K_h - f| \le \sqrt{\frac{1}{2\pi} \int_{|t|>T/h} \phi^2(t)\, dt} .$$

Thus, for densities on [0,1], the bias is directly related to a tail integral of $\phi^2$. As an example of how one can obtain bounds that are explicit in $h$, note that for $\alpha>0$,

$$\int_0^1 |f*K_h - f| \le \sqrt{\frac{1}{2\pi}\left(\frac{h}{T}\right)^{\alpha} \int |t|^{\alpha}\phi^2(t)\, dt} .$$

It should be observed that there are many densities on [0,1] for which $\int |t|^{\alpha}\phi^2(t)\, dt < \infty$ for all $\alpha>0$ (see exercise 7.11). For these densities, the bias tends to zero faster than any polynomial rate in $n$. This implies that $\inf_{h>0} E(\int |f_n - f|) = o(n^{\epsilon}/\sqrt{n})$ for any $\epsilon>0$.

Most ultra-smooth densities, such as the normal density, have infinite support. The reader is urged to look at exercise 7.12 to get an idea of how to cope with the infinite support, since the Cauchy-Schwarz inequality is no longer useful.

Finally, for the ultimate in smoothness, consider a density $f$ with bounded spectrum (i.e, the characteristic function $\phi$ vanishes off $[-S,S]$ for some finite constant $S$). For $h \le T/S$, the bias is zero. Thus, it is not necessary to let $h$ tend to 0 with $n$. In fact, since the variation is roughly speaking a constant over $\sqrt{nh}$, it is best to choose $h$ as large as possible, i.e. $h = T/S$. For a flattop superkernel with $\int \psi^2 < \infty$ and constant $h \le T/S$, we have

$$E(\int |f_n - f|) \le \frac{\int \sqrt{f*(K^2)_h}}{\sqrt{nh}},$$

which decreases as $1/\sqrt{n}$ when $\int \sqrt{f*(K^2)_h} < \infty$. Employing Theorem 7.4, the numerator is further bounded by

$$\int \sqrt{f} + \sqrt{2\pi}\left(4\sigma^2 + 2h^2 \int x^2K^2 / \int K^2\right)^{\frac{1}{4}} .$$

In other words, it is finite if the variance of $f$, $\sigma^2$, and $\int (1+x^2)K^2$ are both finite.

## 7.12. EXERCISES.

7.1. Show that if $\psi$ is the characteristic function of a symmetric function $K$, where $\psi$ has a continuous absolutely integrable derivative of order $s$, then

$$|K(x)| \leq \frac{\int |\psi^{(s)}|}{2\pi |x|^s}$$

where $s$ is a nonnegative integer.

7.2. Consider the "characteristic functions"

$$\psi(t) = \begin{cases} 1 & (|t| \leq 1) \\ \dfrac{1}{|t|} & (|t| > 1) \end{cases}$$

and

$$\phi(t) = \begin{cases} 1 & (|t| \leq 1) \\ 1 - e^{\frac{1}{(|t|-1)^2}} & (|t| > 1) \end{cases}$$

For $\phi$ and $\psi$, prove or disprove that they are flattop superkernels. Hint: either derive the explicit form of $K$, or obtain information about the rate of decrease of $K(x)$ as $|x| \to \infty$.

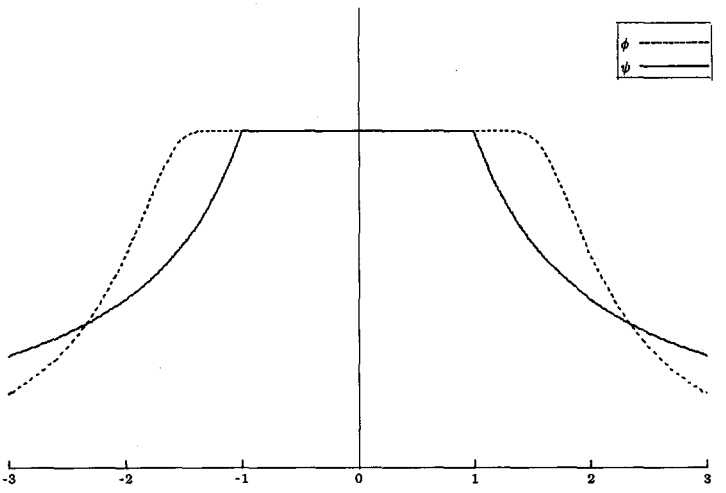


**Figure 7.8.**
Are these characteristic functions of flattop superkernels?

7.3. Prove that there does not exist an unbiased kernel estimate (for any $K$ with $\int K = 1$, $\int |K| < \infty$) if one wants to estimate a density $f$ whose charac-

eristic function is nonzero for all $t$.

how that there does not exist an unbiased kernel estimate for any density hen the kernel $K$ is restricted to be nonnegative.

ind densities $f$, $g$ on the real line such that

$$\int |x| f(x)\,dx < \infty, \int \sqrt{f} = \infty,$$

$$\int |x| g(x)\,dx = \infty, \int \sqrt{g} < \infty.$$

onstruct one density $f$ for which $\int \sqrt{f} < \infty$, yet $\int \sqrt{f * (K^2)_h} = \infty$ for all $>0$ and all bounded kernels $K$ with compact support.

et $f_n$ be a kernel estimate with smoothing factor $h$ satisfying $h \to 0$, $h \to \infty$ as $n \to \infty$. Assume that $K$ is an arbitrary kernel, and that $\sqrt{f} = \infty$. Prove that

$$\liminf_{n \to \infty} \sqrt{nh}\, E\left(\int |f_n - f * K_h|\right) = \infty.$$

ssume that $K$ is a kernel for which $\int (1+|x|^{1+\epsilon}) K^2 < \infty$ and that $f$ is a ensity for which $\int |x|^{1+\epsilon} f(x)\,dx < \infty$ for some $\epsilon > 0$. Show that heorem 7.4 remains valid, i.e.

$$E\left(\int |f_n - f * K_h|\right) \le (1+o(1)) \frac{\sqrt{\int K^2} \int \sqrt{f}}{\sqrt{nh}}$$

f $h \to 0$ as $n \to \infty$.

With a nonnegative order 2 kernel, the best rate of convergence of $E\left(\int |f_n - f|\right)$ achievable with the kernel estimate tends to 0 as $n^{-2/5}$, provided that $D_2^*(f) < \infty$ and $\int \sqrt{f} < \infty$. Give as accurate a statement as possible about the rate of decrease (in $n$) of

$$\inf_{h>0} E\left(\int |f_n - f|\right)$$

where $f$ is one of the following densities:

1. The beta $(a, a)$ density $f(x) = \frac{(2a-1)!}{(a-1)!(a-1)!} (x(1-x))^{a-1}$, $x \in (0,1)$, where $0 < a \le 1$ is a parameter. For $a = 1$ (the uniform density), $D_2^*(f) = \infty$, but $D_1^*(f) < \infty$; for $a < 1$, even $D_1^*(f) = \infty$.

3. The $t_a$ density $f(x) = C_a (1 + x^2/a)^{-\frac{a+1}{2}}$, $x \in R$, where $0 < a \le 1$ is a parameter and $C_a$ is a normalization constant depending upon $a$ only. Note that for all the values of $a$, $\int \sqrt{f} = \infty$, but the size of the tail depends very much on $a$.

For the second family, the tail (and thus the variation) dominates, while for the first family, the oscillation (and thus the bias) is the main contributor to the $L_1$ error.

7.10. Consider a log concave density (log $f$ is concave) with a mode at 0 (log-concave densities are unimodal). Prove or disprove: it is possible to find $h=h_n$ such that $E(\int |f_n-f|)=O(n^{-1/3})$ as $n\to\infty$, where $f_n$ is the kernel estimate with fixed nonnegative order 2 kernel $K$.

7.11. Construct a density $f$ on [0,1] for which $\int |t|^{\alpha}\phi^2(t)\,dt<\infty$ for all $\alpha>0$, where $\phi$ is the characteristic function for $f$.

7.12. With the trapezoidal flattop kernel $(2-|t|)_+-(1-|t|)_+$, how would you choose $h$ for a normal density $f$? What rate is achievable for $E(\int |f_n-f|)$? Hint: use Theorem 7.6 for an upper bound on the variation. Handle the bias term either via an inequality in the spirit of Carlson's inequality, or via an integral split in which Cauchy-Schwarz is used on a compact (but growing) central interval, and crude bounds are used for (shrinking) tail intervals. Recall that if $X$ is normally distributed, then $P(X>u)\sim u^{-1}f(u)$ as $u\to\infty$, where $f$ is the normal density.

7.13. Let $K$ be a nonnegative kernel, and let $\mathbf{F}$ be the class of uniformly continuous densities with support on [0,1] and modulus of continuity $\sup_{x,|y|\le\delta}|f(x+y)-f(x)|$ bounded by a function $\omega(\delta)\to 0$ as $\delta\downarrow 0$. Show that

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

Show that for all classes $W(s,\alpha,C)$, such a function $\omega(\delta)$ can be found.

7.14. Unbiased density estimates can exist for certain classes of densities. For example, Kolmogorov, and later Basu (1964) have shown that for the normal family with unknown mean $\mu$ and variance $\sigma^2$, the following density is an unbiased estimate at all $x$:

$$f_n(x)=\frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n-2}{2})\sqrt{\pi(n-1)}\hat{\sigma}}\left(1-\frac{1}{(n-1)\hat{\sigma}^2}(x-\hat{\mu})^2\right)_+^{\frac{n-4}{2}} .$$

Here $\hat{\mu}$ and $\hat{\sigma}$ are the standard sample-based estimates of $\mu$ and $\sigma$. Show that the given estimate is indeed unbiased at all $x$, and prove that $E(\int |f_n-f|)=O(1/\sqrt{n})$ uniformly in $\mu$ and $\sigma^2$. For other examples and more theoretical background on unbiased estimation, see Lumelskii and Sapozhnikov (1969), Wertz (1975), Guttmann and Wertz (1976) and Seheult and Quesenberry (1971), and the references found there.

7.15. Is every Lipschitz density $f$ absolutely continuous (see definition at beginning of chapter VII)? Prove your answer.

7.16. Show that if $K$ is a kernel with order greater than $s$, and $D_s^*(f)<\infty$, then $\int |f*K_h-f|=o(h^s)$ as $h\downarrow 0$. Note: when $f$ has $s-1$ absolutely continuous derivatives, the result is implicit in Theorem 7.2.

7.17. Show that for **any** kernel estimate $f_n$, $\inf_{f,K,h} E(\int |f_n-f|)\ge c/\sqrt{n}$ for some positive constant $c$ ($c=1/\sqrt{526}$ will do). Hint: use the realtionship

with characteristic functions, split bias and variation terms, and employ the inequality $E(|S_n|) \geq E(|X|)\sqrt{n}/8$, which is valid for all sums $S_n$ of $n$ iid zero mean random variables distributed as $X$.

8. This exercise should provide you with some idea about the smoothness of $f$ when $f$ has bounded spectrum. Show that when a density $f$ has bounded spectrum, it has a Taylor series expansion about the origin which is uniformly convergent in any bounded interval. Relate the coefficients of the series to the characteristic function $\phi$ of $f$.

9. The variation of the kernel estimate can be larger than $c/\sqrt{nh}$ when $\int\sqrt{f} = \infty$. Show that the inequality of Theorem 7.3 can be generalized as follows: if $|K| \leq K_{\max}$, then

$$E\left(\int |f_n - f*K_h|\right) \leq 2^p (K_{\max})^{\frac{1-p}{2}} \frac{\int (f*|K|_h)^{\frac{1+p}{2}}}{(nh)^{\frac{1-p}{2}}},$$

where $p \in (0,1)$. Find sufficient conditions on $f$ and $K$ for the upper bound to be finite when $\int f^{\frac{1+p}{2}} < \infty$.

10. Let $f$ be a very smooth density, in the sense that $|\phi''(t)| \leq C^2 e^{-|t|/D}$ where $\phi$ is the characteristic function for $f$ and $\phi''$ is assumed to exist at all $t$. In other words, the characteristic function decreases at an exponential rate or faster. Design a kernel estimate for which

$$E\left(\int |f_n - f|\right) \leq (\gamma + o(1)) \sqrt{\frac{CD \log(n)}{n}}$$

uniformly over this class, for some universal constant $\gamma > 0$. Give explicit formulas for $h$ and $K$ as a function of $C$ and $D$. Compute acceptable values for $\gamma$ and $o(1)$ in the upper bound. Interpret $C$ and $D$, and the "difficulty factor $\sqrt{CD}$", which depends only upon the shape of $f$. Consider the plane of densities with coordinates $C$ and $D$ (similar to figure 7.7), and show at least five densities in this plane, including the normal density.

# Chapter Eight
# A CASE STUDY: MONOTONE DENSITIES ON [0,1]

## 8.1. SCOPE OF THIS CHAPTER.

We study the class $\mathbf{F}=M_B$ of bounded monotone (nonincreasing) densities on [0,1], where $B$ is an upper bound for the value of the density at the peak (the origin).

The estimators being studied include

A. Grenander's estimate (also called the SCM or smallest concave majorant estimate).

B. The kernel estimate.

C. The histogram estimate.

D. Birge's modified histogram estimate.

The class $\mathbf{F}$ includes many smooth densities, for which a modified kernel estimate can in fact achieve extremely good rates of convergence depending upon the smoothness of the density (a modification is needed to treat the discontinuity at the origin; see further on). But smoothness is not what we are interested in here. The monotonicity itself is the interesting ingredient. How can we incorporate that knowledge in our estimate, and what can we hope to achieve? To fix a specific goal, we will begin with the computation of a minimax lower bound in Theorem 8.1. We will see that the kernel and histogram estimates are not bad, but they are not minimax-optimal because none exploits the monotonicity very well. Birge's modified histogram estimate is minimax-optimal, and provides us with a splendid example of how estimates can be tailor made for certain classes of densities.

The key references are Grenander (1956,1981), who applies the maximum likelihood principle in the definition of an estimate, Groeneboom (1983), who provides a thorough analysis of Grenander's estimate, and Birge (1983,1984).

## 8.2. THE MINIMAX LOWER BOUND.

**Theorem 8.1.**

$$\inf_{f_n} \sup_{f \in M_B} E\left(\int |f_n - f|\right) \geq 1 ,$$

and

$$\inf_{f_n} \sup_{f \in M_B} E\left(\int |f_n - f|\right) \geq \frac{1}{8 + 4\left(\frac{4S}{n}\right)^{\frac{1}{3}}} \left(\frac{4S}{n}\right)^{\frac{1}{3}} \quad (3 \leq S \leq 0.054n)$$

where $S = \log(1+B)$.

If we were to construct a subclass of monotone densities based upon a partition of [0,1] into intervals of equal length, then we would obtain at best a lower bound of the form $c/n^{1/3}$ where $c$ is a constant not depending upon $B$ (for an explicit computation of this type, see Devroye and Gyorfi, 1985). To obtain a bound with the correct dependence upon $B$ and $n$, not just $n$, Birge (1983) observed that it is necessary to consider unequal intervals. Theorem 8.1 and the proof given below are essentially his. We took the liberty to shorten the proof a bit; the price paid for it is a slightly worse constant in the bound.

**Proof of Theorem 8.1.**

Let $0 < \epsilon \leq 0.6$ be a real number, and let $r \geq 1$ be an integer, both to be picked later. Then partition [0,1] into $r$ intervals

$$A_i = [x_{i-1}, x_i) , \ i = 1, \ldots, r ,$$

where

$$x_i = \frac{(1+\epsilon)^i - 1}{(1+\epsilon)^r - 1} , \ i = 0, 1, \ldots, r .$$

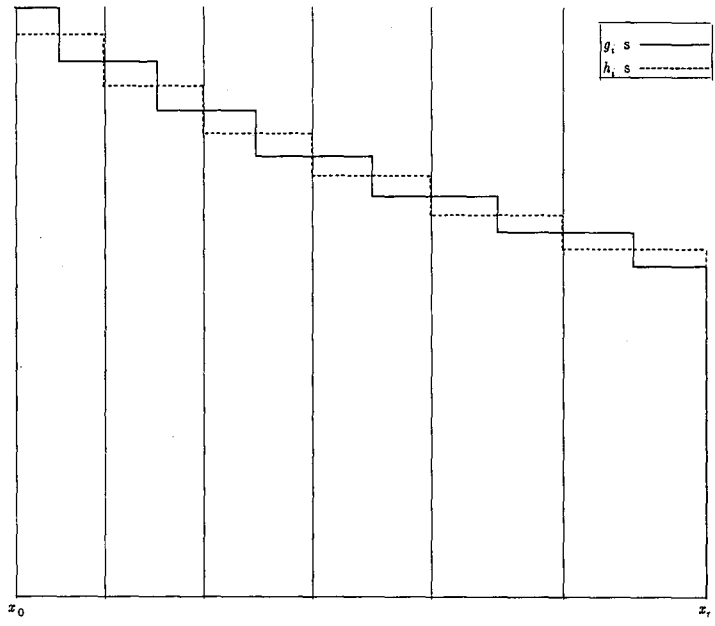


**Figure 8.1.**
Partition of [0,1] needed to construct subclass of $\mathbf{F}$.

On $A_i$, $f_\theta$ is either $g_i$ or $h_i$, depending upon whether $\theta_i$ is zero or one. Recall that $\theta=0.\theta_1\theta_2\cdots$ is a binary expansion of a number $\theta\in[0,1]$. The functions $g_i$ and $h_i$ shown in figure 8.1 are staircase functions, where

$$h_i = \frac{\lambda(1+\frac{\epsilon}{2})}{(1+\epsilon)^i},$$

$$g_i = \frac{\lambda(1+\epsilon)}{(1+\epsilon)^i}, \text{ or } \frac{\lambda}{(1+\epsilon)^i},$$

where

$$\lambda = \frac{1+\epsilon}{r\,\epsilon(1+\frac{\epsilon}{2})}\left((1+\epsilon)^r - 1\right).$$

Observe that $\int_{A_i} h_i = \int_{A_i} g_i = \frac{1}{r}$ since the length of $A_i$ is

$$x_i - x_{i-1} = \frac{(1+\epsilon)^i - (1+\epsilon)^{i-1}}{(1+\epsilon)^r - 1} = \frac{\epsilon(1+\epsilon)^{i-1}}{(1+\epsilon)^r - 1} = \frac{1}{rh_i} .$$

Let us now compute $\alpha$ and $\beta$ for use in Assouad's bound (see Theorem 5.2):

$$\int_{A_i} | h_i - g_i | = \frac{\lambda \frac{\epsilon}{2}}{(1+\epsilon)^i} \times \text{length } A_i$$

$$= \frac{\lambda \epsilon^2}{2(1+\epsilon)((1+\epsilon)^r - 1)} = \frac{\epsilon}{r(2+\epsilon)} \stackrel{\Delta}{=} \alpha .$$

Furthermore,

$$\int_{A_i} \left( \sqrt{h_i} - \sqrt{g_i} \right)^2 \leq \frac{\epsilon^2}{16\, r} \left(1 + \frac{\epsilon}{4}\right)$$

$$\stackrel{\Delta}{=} 2 - 2\beta .$$

The last inequality is most easily shown by using the fact that $\sqrt{1+x} \geq 1 + \frac{x}{2} - \frac{x^2}{8}$ for $x > 0$:

$$\int_{A_i} \left( \sqrt{h_i} - \sqrt{g_i} \right)^2$$

$$= \frac{\epsilon(1+\epsilon)^{i-1}}{(1+\epsilon)^r - 1} \frac{\lambda}{(1+\epsilon)^i} \left( \frac{1}{2} \left( \sqrt{1 + \frac{\epsilon}{2}} - \sqrt{1+\epsilon} \right)^2 + \frac{1}{2} \left( 1 + \sqrt{1 + \frac{\epsilon}{2}} \right)^2 \right)$$

$$= \frac{\epsilon}{1+\epsilon} \frac{1+\epsilon}{r\,\epsilon\left(1+\frac{\epsilon}{2}\right)} \frac{1}{2} \left( 4 + 2\epsilon - 2\sqrt{1 + \frac{\epsilon}{2}} \left(1 + \sqrt{1+\epsilon}\right) \right)$$

$$\leq \frac{1}{r(2+\epsilon)} \left( 4 + 2\epsilon - 2\left(1 + \frac{\epsilon}{4} - \frac{\epsilon^2}{32}\right)\left(2 + \frac{\epsilon}{2} - \frac{\epsilon^2}{8}\right) \right)$$

$$= \frac{1}{r(2+\epsilon)} \left( \frac{\epsilon^2}{8} + \frac{3\epsilon^3}{32} - \frac{\epsilon^4}{128} \right)$$

$$= \frac{\epsilon^2 \left(1 + \frac{3\epsilon}{4} - \frac{\epsilon^2}{16}\right)}{16 r \left(1 + \frac{\epsilon}{2}\right)} \leq \frac{\epsilon^2}{16\, r} \left(1 + \frac{\epsilon}{4}\right) .$$

We can plug everything back into Assouad's lower bound

$$\frac{r\,\alpha}{2} \left( 1 - \sqrt{2 - 2\beta^n} \right) \geq \frac{r\,\alpha}{2} \left( 1 - \sqrt{2n\,(1-\beta)} \right)$$

$$= \frac{\epsilon}{2(2+\epsilon)} \left( 1 - \sqrt{\frac{n\,\epsilon^2}{16\, r} \left(1 + \frac{\epsilon}{4}\right)} \right) .$$

We can make the square root less than $\frac{1}{2}$, if we take

$$r = \left\lceil \frac{n\epsilon^2}{4}(1+\frac{\epsilon}{4}) \right\rceil .$$

We are able to conclude that

$$m(n,\mathbf{F}) \geq \frac{\epsilon}{8(1+\frac{\epsilon}{2})} .$$

This expression, or $\frac{\epsilon}{8}$, should be maximized with respect to $\epsilon$, subject to the only constraint we have not verified yet, i.e. that we indeed have a subclass of $\mathbf{F}$, i.e. that $g_1 = \lambda \leq B$. It is important to see at this stage how one should choose $\epsilon$. Observe that $\lambda \approx e^{r\epsilon}$. This is at most $B$ if $r\epsilon \leq S$, roughly speaking. Substituting $r$ by its approximate value, $n\epsilon^2/4$, we obtain that $n\epsilon^3/4 \leq S$. This is why the value

$$\epsilon = \left(\frac{4S}{n}\right)^{\frac{1}{3}}$$

is approximately optimal. With this value, we obtain the lower bound that we wanted to verify. It suffices to show now that $\lambda \leq B$ with this choice of $\epsilon$.

First four small observations: (1) $S < 0.054n$ implies that $\epsilon \leq 0.6$; (2) $r < 1+\frac{S}{\epsilon}(1+\frac{\epsilon}{4})$; (3) $\frac{1}{x}((1+\epsilon)^x - 1)$ is nondecreasing in $x$ for $x > 0$; (4) $(1+\frac{\epsilon}{4})\log(1+\epsilon) \leq \epsilon$. This is used in the following chain of inequalities:

$$\lambda = \frac{1+\epsilon}{r\,\epsilon(1+\frac{\epsilon}{2})}\left((1+\epsilon)^r - 1\right)$$

$$< \frac{(1+\epsilon)e^{\frac{S}{\epsilon}(1+\frac{\epsilon}{4})\log(1+\epsilon)} - 1}{\frac{\epsilon(1+\frac{\epsilon}{2})}{1+\epsilon} + \frac{S(1+\frac{\epsilon}{2})(1+\frac{\epsilon}{4})}{1+\epsilon}}$$

$$< \frac{(1+\epsilon)e^S - 1}{S+\epsilon}\,\frac{1+\epsilon}{1+\frac{\epsilon}{2}}$$

$$< e^S - 1 .$$

Only the last inequality requires explicit verification. This boils down to verifying

$$e^S(1+\epsilon+\frac{\epsilon^2}{2} - S - \frac{S\epsilon}{2}) < 1 - \epsilon - S - \frac{S\epsilon}{2} .$$

For $S \geq 3$, we have $e^S \geq 10$, and $S-1+S\epsilon/2 \geq 2 \geq \epsilon+\epsilon^2/2$. So, the left-hand side of the last inequality is at most equal to

$$10(1+\epsilon+\frac{\epsilon^2}{2}-S-\frac{S\epsilon}{2}) .$$

Thus, we need to check that $9(S-1+S\epsilon/2) > \epsilon+10(\epsilon+\epsilon^2/2)$. This is immediate from the fact that $S \geq 3$ and that $\epsilon \leq 0.6$. ■

## 8.3. GRENANDER'S ESTIMATE.

In 1956, Grenander studied an estimator of a monotone density which did not require knowledge of the support of $f$, and which itself was a monotone density. The latter restriction allowed him to apply the maximum likelihood principle very elegantly. In this section we will study some of the properties of Grenander's estimate, also referred to below as the **MLE**.

Grenander's estimate $f_n$ is defined as a monotone density with the property that

$$\prod_{i=1}^{n} f_n(X_i)$$

is maximal. We need a few lemmas that will enable us to say more about the shape of this estimate.

**Lemma 8.1.**

The MLE $f_n$ is a step function with breakpoints (jumps) at the order statistics $X_{(i)}, 1 \leq i \leq n$.

The general form of an MLE is shown below:

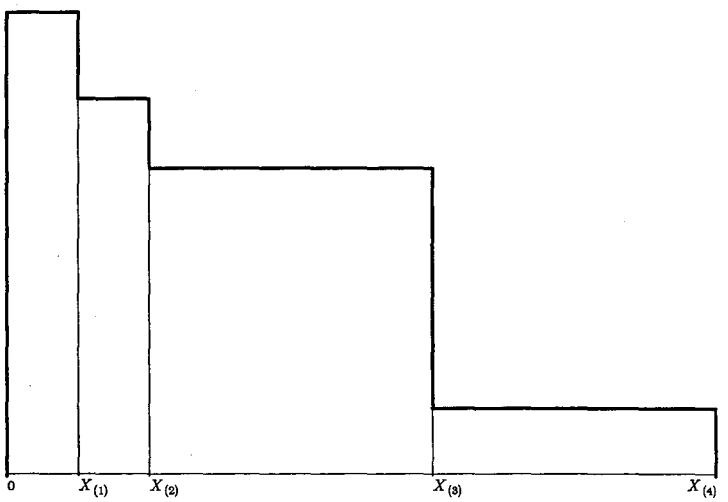


**Figure 8.2.**
General form of an MLE.

**Proof of Lemma 8.1.**

The log-likelihood of the sample with order statistics $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ is

$$L(f) \stackrel{\Delta}{=} \sum_{i=1}^{n} \log(f(X_{(i)})) .$$

Define

$$f^*(x) = \begin{cases} 0 & (x \leq 0) \\ Cf(X_{(i)}) & (X_{(i-1)} < x \leq X_{(i)}) \\ 0 & (x > X_{(n)}) \end{cases}$$

where $C$ is a normalizing constant. Observe that

$$L(f^*) = n \log(C) + L(f) \geq L(f)$$

since $C \geq 1$ (this follows from the fact that $\sum (X_{(i)} - X_{(i-1)}) f(X_{(i)}) \leq \int_0^{X_{(n)}} f \leq 1$). Thus, for every density $f_n$, there exists a step function $f_n^*$ with breakpoints at

the order statistics, for which $L(f_n^*) \geq L(f_n)$. ■

We will also need some fact about entropies of discrete distributions:

**Lemma 8.2.**

Let $p_1, \ldots, p_n$ and $q_1, \ldots, q_n$ be probability vectors with positive components. Then

$$\sum_{i=1}^{n} q_i \log(\frac{q_i}{p_i}) \geq 0 .$$

**Proof of Lemma 8.2.**

We can either base the proof on the positivity of Kullback-Leibler numbers (i.e., $\int f \log(\frac{f}{g}) \geq 0$ for all densities $f, g$), or prove it directly. Recalling that the function $u \log(u)$ is convex in $u$, we have, by Jensen's inequality,

$$\sum_{i=1}^{n} p_i \frac{q_i}{p_i} \log(\frac{q_i}{p_i}) \geq \sum_{i=1}^{n} p_i \frac{q_i}{p_i} \log(\sum_{i=1}^{n} p_i \frac{q_i}{p_i}) = 1 \log(1) = 0 . \blacksquare$$

From Lemma 8.1, we retain that the form of the MLE is that of a data-dependent histogram. The third lemma applies to all histogram estimates, and is presented in this setting. It states that once we have settled on breakpoints, the MLE is completely specified.

**Lemma 8.3.**

Consider a partition $A_1, \ldots, A_k$ of a compact set $A$, and a histogram density estimate $f_n$ taking the value $g_i$ on $A_i$, subject to the normalization $\sum_i g_i \lambda(A_i)=1$. Then, the maximum over all these histogram estimates of the likelihood product

$$\prod_{i=1}^{n} f_n(X_i)$$

is reached for the histogram estimate with

$$g_i = \frac{\mu_n(A_i)}{\lambda(A_i)}$$

where $\mu_n$ is the empirical measure for the data.

**Proof of Lemma 8.3.**

Define $\Delta_i = \lambda(A_i)$, and $C_i =$ cardinality of $A_i$, or $n\,\mu_n(A_i)$. Observe that for any vector $g_1, \ldots, g_k$,

$$\prod_{j=1}^{n} f_n(X_i) = \prod_{i=1}^{k} g_i^{C_i}$$

$$= \prod_{i=1}^{k} \left(\frac{g_i\ n\ \Delta_i}{C_i}\right)^{C_i} \prod_{i=1}^{k} \left(\frac{C_i}{n\ \Delta_i}\right)^{C_i}$$

$$\leq \left( \frac{\sum_{i=1}^{k} g_i\ n\ \Delta_i}{\sum_{i=1}^{k} C_i} \right)^{n} \prod_{i=1}^{k} \left(\frac{C_i}{n\ \Delta_i}\right)^{C_i}$$

(arithmetic-geometric mean inequality)

$$= \prod_{i=1}^{k} \left(\frac{C_i}{n\ \Delta_i}\right)^{C_i}.$$

This concludes the proof of the lemma. ■

When Lemma 8.3 is applied with $A_i = (X_{(i-1)}, X_{(i)}]$ (with $0 = X_{(0)}$ by convention, and $1 \leq i \leq n$), then it is easily seen that among all step function densities with breakpoints at the order statistics, the likelihood product is maximized if we take a density which, on $(X_{(i-1)}, X_{(i)}]$, takes the value $1/(n\,(X_{(i)} - X_{(i-1)}))$ (since the empirical measure of each interval is precisely one).

We can now formulate

**Theorem 8.2. Grenander's theorem.**

The monotone density on $[0,\infty)$ for which the likelihood product is maximal is the density whose distribution function is the smallest concave majorant of the empirical distribution function.

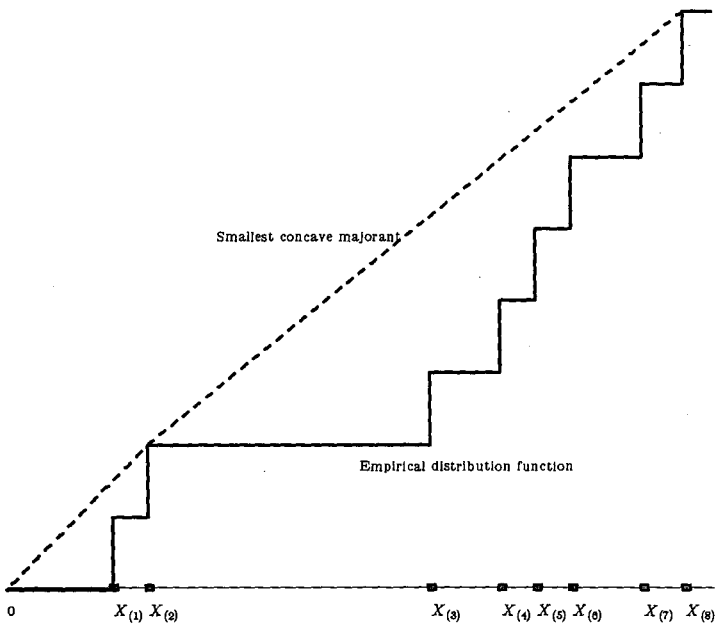


**Figure 8.3.**
The empirical distribution function and its smallest concave majorant are shown.

**Proof of Theorem 8.2.**

By Lemma 8.1, it suffices to consider only monotone histograms with breakpoints at the order statistics. Consider just such a density, and let its height be $g_i$ on the interval $A_i=(X_{(i-1)},X_{(i)}]$. Let $\Delta_i$ be the length of $A_i$. Consider a partition of $1,\ldots,n$ into intervals of indices, $I_1,I_2,\ldots,I_k$, and define

$$p_i = \sum_{j\in I_i} g_j \Delta_j ,$$

$$q_i = \frac{1}{n}\text{Cardinality of } \bigcup_{j\in I_i} A_j ,$$

$$h_j = \left(\frac{q_i}{p_i}\right) g_j \ , \ j\in I_i ,$$

where $i=1,2,\ldots,k$. Note that the $h_j$'s define another histogram density estimate, and that it has the following properties:

A. It integrates to one, as

$$\sum_j h_j \Delta_j = \sum_{i=1}^{k} \left(\frac{q_i}{p_i}\right) \sum_{j\in I_i} g_j \Delta_j = 1 .$$

B. It has a better likelihood product than the $g_i$'s, since

$$\prod_{i=1}^{k} \prod_{j\in I_i} h_j = \prod_{i=1}^{k} \left(\frac{q_i}{p_i}\right)^{nq_i} \prod_{j\in I_i} g_j$$

$$\geq \prod_{i=1}^{k} \prod_{j\in I_i} g_j$$

by Lemma 8.2.

This improvement is applicable to any histogram estimate. We can pick any partition we like. In particular, we can partition the indices by the breakpoints of the smallest concave majorant of the empirical distribution function. If we do that, then the improvement $h_j$ agrees with the empirical distribution function at these points: indeed,

$$\sum_{j\in I_i} h_j \Delta_j = \frac{q_i}{p_i} \sum_{j\in I_i} g_j \Delta_j$$

$$= q_i = \frac{1}{n}\text{Cardinality of } \bigcup_{j\in I_i} A_j .$$

Furthermore, since the $g_j$'s are nonincreasing, the $h_j$'s are nonincreasing within one set $I_i$ (they may not be across index sets). Finally, by our choice of breakpoints, we have that

$$\frac{\Delta_{i+1}+\cdots+\Delta_{i+m}}{m} \geq \frac{\Delta_{i+1}+\cdots+\Delta_{i+k}}{k}$$

for $1\leq m\leq k$, as long as we stay in one index set. Having started from

$g_1, \ldots, g_n$, and improved it to $h_1, \ldots, h_n$, we now present a further improvement to $l_1, \ldots, l_n$, which has the property that it is independent of the original choice of $g_i$'s. The improvement coincides with Grenander's estimate, and is an MLE. All $l_j$'s for $j \in I_i$ are equal to

$$\frac{\sum_{j \in I_i} h_j \Delta_j}{\sum_{j \in I_i} \Delta_j} = \frac{q_i}{\sum_{j \in I_i} \Delta_j},$$

i.e. they agree with Grenander's estimate. It suffices to show that we have a likelihood product improvement for every $I_i$. To see this, we need to show that

$$\prod_{j \in I_i} h_j \leq \prod_{j \in I_i} l_j$$

for all $i$. By the arithmetic-geometric mean inequality, we have

$$\left( \prod_{j \in I_i} h_j \right)^{\frac{1}{nq_i}} \leq \frac{1}{nq_i} \sum_{j \in I_i} h_j$$

which in turn does not exceed

$$\frac{\sum_{j \in I_i} h_j \Delta_j}{\sum_{j \in I_i} \Delta_j} \stackrel{\Delta}{=} l_j$$

by virtue of an interesting association inequality explored in exercise 8.1. This concludes the proof of Grenander's theorem. ■

This is not the proper forum for exploring all the properties of Grenander's estimate. Clearly, the fact that it is completely defined, without smoothing parameters and the like, is appealing. Let us merely point out that the estimate is consistent for monotone densities and inconsistent for all other densities (Theorem 8.3). Furthermore, for smooth monotone densities, the individual rate of convergence is $n^{-1/3}$ (Theorem 8.4). See also exercise 8.2 about the minimax error on small subclasses of monotone densities.

**Theorem 8.3.**

Let $f_n$ be Grenander's estimate. Then $\int | f_n - f | \to 0$ almost surely if and only if $f$ is monotone.

**Proof of Theorem 8.3.**

If $f$ is not monotone, then

$$\int | f_n - f | \geq \inf_{g \text{ monotone}} \int | g - f | > 0 .$$

When $f$ is monotone, $F$ is its distribution function, and $F_n$ and $\hat{F}_n$ are the empirical distribution function and the smallest concave majorant of the empirical distribution function respectively, then

$$f_n(x) \leq \frac{1}{\delta}(\hat{F}_n(x) - \hat{F}_n(x-\delta)) \quad \text{(any } \delta > 0\text{; by concavity)}$$

$$\leq \frac{2}{\delta} \sup_y | \hat{F}_n(y) - F(y) | + \frac{1}{\delta} \int_{x-\delta}^{x} f - f(x) + f(x) .$$

The first term on the right-hand side is not greater than $\frac{2}{\delta} \sup_y | F_n(y) - F(y) |$ by the concavity of $F$. It tends to 0 almost surely by virtue of the Glivenko-Cantelli theorem. The Lebesgue density theorem implies that for almost all $x$, we can find $\delta$ so small that the second and third terms taken together are as small as desired. We can argue symmetrically on

$$f_n(x) \geq \frac{1}{\delta}(\hat{F}_n(x+\delta) - \hat{F}_n(x)) .$$

This shows that $f_n \to f$ almost surely for almost all $x$. Since $\int f_n = 1$, we can apply Glick's theorem and conclude that $\int | f_n - f | \to 0$ almost surely. ∎

**Theorem 8.4.**

Assume that $f$ has two bounded continuous derivatives on $(0,\infty)$, and that $f' < 0$ on $(0,1)$. Then

$$n^{\frac{1}{3}} E\left(\int | f_n - f |\right) \to c \int \left(\frac{1}{2} f | f' |\right)^{\frac{1}{3}}$$

where $f_n$ is Grenander's estimate and $c \approx 0.82$ is a universal constant (Groeneboom, 1983).

Theorem 8.4 is not shown here. The rate is that predicted by the minimax lower bounds. In that respect, Theorem 8.4 states that individually, we observe the same rate of convergence as we can hope to observe uniformly over $\mathbf{F}$, even

for the smoothest densities in the class. But since we can construct much faster converging kernel estimates (properly modified to handle the discontinuity at the origin, see next section) for densities dealt with in Theorem 8.4, Theorem 8.4 should be regarded as a bad news message for Grenander's estimate. There is of course still hope that the estimate is minimax-optimal: for one thing, the constant in Theorem 8.4 has the right dependence upon $S$ (see Lemma 8.4 below). See also exercise 8.2.

**Lemma 8.4.**

Let $f \in \mathbf{F}$ be absolutely continuous. Then

A. $\int \left( f \mid f' \mid \right)^{\frac{1}{3}} \leq 1 + (\log B)^{\frac{1}{3}} \leq 1 + S^{\frac{1}{3}}$.

B. $\inf_{f \in \mathbf{F},\, f \text{ absolutely continuous}} \int \left( f \mid f' \mid \right)^{\frac{1}{3}} = 0$.

**Proof of Lemma 8.4.**

Let $u$ be a number in $[0,1]$ with the property that $f(u) = 1$ (see figure 8.4).

$$\int \left( f \mid f' \mid \right)^{\frac{1}{3}} \leq \int_0^u \int \left( f \mid f' \mid \right)^{\frac{1}{3}} + \int_u^1 \mid f' \mid^{\frac{1}{3}}$$

$$\leq \int_0^u f \left( \frac{\mid f' \mid}{f^2} \right)^{\frac{1}{3}} + f(u)^{\frac{1}{3}}$$

$$\leq \left( \int_0^u \frac{\mid f' \mid}{f^2} \right)^{\frac{1}{3}} + 1$$

$$= \left( \int_0^u -d(\log f) \right)^{\frac{1}{3}} + 1$$

$$= \left( \log B \right)^{\frac{1}{3}} + 1 .$$

The second half of Lemma 8.4 is trivial.

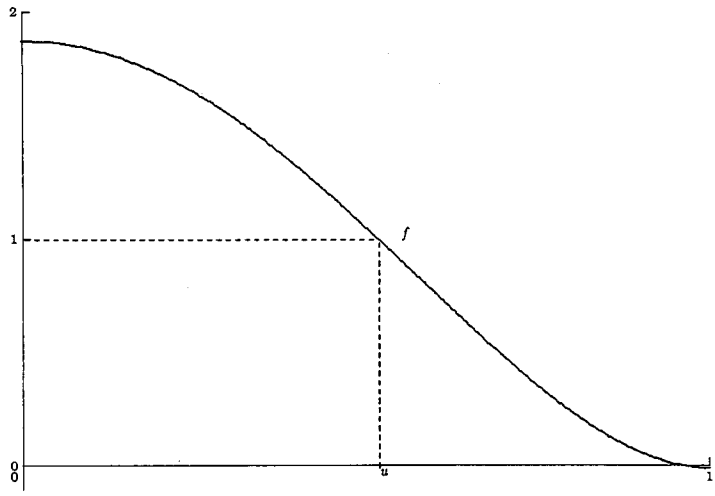


**Figure 8.4.**
A monotone density.

■

## 8.4. THE KERNEL ESTIMATE.

The **symmetrized kernel estimate** is defined by

$$f_n(x) = f_n^*(x) + f_n^*(-x), \quad x > 0,$$

where $f_n^*$ is the standard kernel estimate with a symmetric kernel $K$ based upon $S_1X_1, \ldots, S_nX_n$, where $S_1, \ldots, S_n$ are iid random signs. The support of $f_n$ is the positive real line. Equivalently, $f_n$ can be obtained from the raw kernel estimate $g_n$ by defining

$$f_n(x) = g_n(x) + g_n(-x), \quad x > 0,$$

i.e. by flipping the part of $g_n$ to the left of the origin over and adding it to the part to the right of the origin, see figure 8.5.

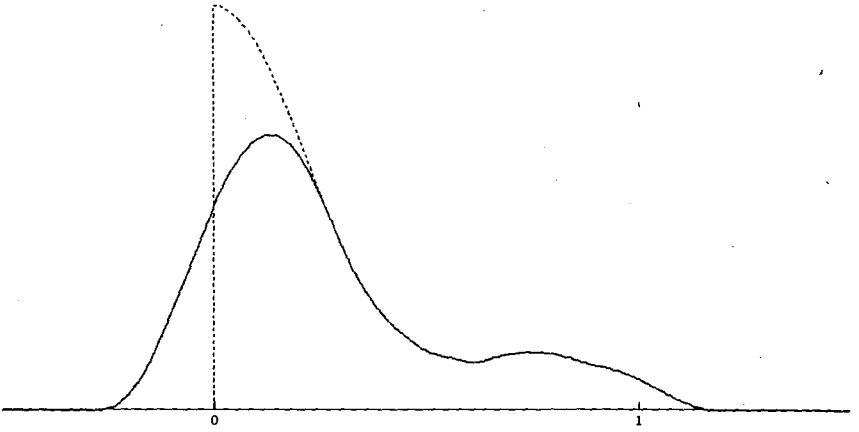


**Figure 8.5.**
A kernel estimate $g_n$ and its symmetrized form $f_n$.

The inequality

$$\int |f_n - f| \leq \int |g_n - f|$$

valid for any $f$ with support on $[0,\infty)$ is of little help to us, since the $f$ to be estimated still has a discontinuity at the origin; it just tells us that flipping the mass around is harmless. What is needed is

**Lemma 8.5.**

Let $f_n$ be the symmetrized kernel estimate based upon a sample drawn from a density $f$ with support on $[0,\infty)$, and let $f_n^*$ be the kernel estimate based upon a sample $S_1 X_1, \ldots, S_n X_n$. Then

$$\int_0^\infty |f_n - f| \leq \int_{-\infty}^\infty |f_n^*(x) - \frac{1}{2} f(|x|)| \, dx .$$

(Observe that $f_n^*$ is the kernel estimate of $f(|x|)/2$, a density with support on the real line.)

**Proof of Lemma 8.5.**

The symmetry in $K$ and the data can be exploited as follows:

$$\int_0^\infty |f_n - f| = \int_0^\infty |f_n^*(x) + f_n^*(-x) - f(x)| \, dx$$

$$\leq \int_0^\infty |f_n^*(x) - \frac{1}{2} f(x)| \, dx + \int_0^\infty |f_n^*(-x) - \frac{1}{2} f(x)| \, dx$$

$$= \int_{-\infty}^\infty |f_n^*(x) - \frac{1}{2} f(|x|)| \, dx \; . \blacksquare$$

The point is that $f(|x|)/2$ does not have a discontinuity at the origin, so that $f_n^*$ can possibly approximate it with errors of the order of $n^{-2/5}$ or better. Here we won't even need to go that far in terms of errors. We will call $f_n$ the **F**-tailored kernel estimate if it is a symmetrized kernel estimate with

$$K(x) = (1 - |x|)_+$$

and

$$h = \left( \frac{12}{nB^2} \right)^{\frac{1}{3}} .$$

The smoothing factor is picked for optimal performance uniformly over **F**:

**Theorem 8.5.**

Let $f_n$ be the **F**-tailored kernel estimate. Then

$$\limsup_{n \to \infty} \sup_{f \in \mathbf{F}} n^{\frac{1}{3}} E\left( \int |f_n - f| \right) \leq (3B)^{\frac{1}{3}} .$$

**Proof of Theorem 8.5.**

By Lemma 8.5, it suffices to consider the symmetric density $g(x) = f(|x|)/2$, and the standard kernel estimate $f_n^*$, based upon a sample of size $n$ drawn from $g$. The quantity $E\left( \int |f_n^* - g| \right)$ is first bounded from above by the bias term plus the variation term. From Theorem 7.1, we recall that the bias does not exceed

$$h \int |x| K \liminf_{a \downarrow 0} \int |(g * \phi_a)'|$$

where $\phi$ is a symmetric unimodal mollifier (so that $g^* \phi_a$ is symmetric unimodal as the convolution of two symmetric unimodal densities). We have

$$\int | (g^* \phi_a)' | = 2g^* \phi_a(0) \leq 2 \sup | g | \int | \phi_a | = B .$$

Thus, the bias does not exceed $Bh \int | x | K$. The variation is handled on the basis of Theorem 7.3. We have

$$E(\int | f_n{}^* - g^* K_h |) \leq \frac{1}{\sqrt{nh}} \int \sqrt{g^* (K^2)_h}$$

$$\leq \sqrt{\int K^2} \left( \int \sqrt{g} + \int \sqrt{| g - g^* Q_h |} \right) \quad (Q \stackrel{\Delta}{=} K^2 / \int K^2)$$

$$\leq \sqrt{\int K^2} \left( \int \sqrt{g} + \sqrt{2+2h} \sqrt{\int | g - g^* Q_h |} \right) \text{ (Cauchy-Schwarz)}$$

$$\leq \sqrt{\int K^2} \left( \sqrt{2} \int \sqrt{f} + \sqrt{2+2h} \sqrt{Bh \int | x | Q^2} \right)$$

so that

$$E(\int | f_n - f |) \leq (1+o(1)) \left( Bh \int | x | K + \sqrt{\frac{2 \int K^2}{nh}} \int \sqrt{f} \right) .$$

The value of $h$ for which the main factor is minimized is

$$h = n^{-\frac{1}{3}} \left( 2 \int K^2 \right)^{\frac{1}{3}} \left( \int \sqrt{f} \right)^{\frac{2}{3}} \left( B \int | x | K \right)^{-\frac{2}{3}} .$$

The corresponding value of the main factor in the upper bound is

$$\left( 2^{-\frac{2}{3}} + 2^{\frac{1}{3}} \right) n^{-\frac{1}{3}} \left( B \int | x | K \; 2 \int K^2 \int^2 \sqrt{f} \right)^{\frac{1}{3}}$$

$$\leq (1+2) \left( \frac{B}{n} \right)^{\frac{1}{3}} 3^{-\frac{2}{3}}$$

$$= \left( \frac{3B}{n} \right)^{\frac{1}{3}} ,$$

where we used the fact that for our $K$, $\int | x | K = \frac{1}{3}$, $\int K^2 = \frac{2}{3}$, and that $\int \sqrt{f} \leq 1$. ■

The dependence upon $n$ in the upper bound of Theorem 8.5 matches that of Birge's lower bound. Unfortunately, the dependence upon $B$ is suboptimal, as a factor of $\log^{1/3} B$ is called for, instead of $B^{1/3}$. This reflects the fact that the kernel estimate does not make good use of the information that the density is

monotone. On the other hand, the kernel estimate is consistent for all densities. This distinguishes it from Grenander's estimate and the minimax-optimal estimate presented in the next section. In particular, since every $L_1$-neighborhood of a monotone density contains many non-monotone densities, Grenander's estimate and the estimate of next section are not robust.

It is not necessary to dwell on the standard histogram estimate with equal-length bins, since the results are comparable to those obtained with the kernel estimate. Instead, we will close this case study by presenting Birge's modified histogram estimate (Birge, 1984), which is minimax optimal for **F**.

## 8.5. BIRGE'S MODIFIED HISTOGRAM ESTIMATE.

In 1984, Birge observed that a minimax-optimal estimate for our class **F** could be constructed merely by considering intervals with geometrically increasing interval widths, and fine-tuning the geometrical rate of increase. Note that the same technique was used in Theorem 8.1 in the construction of a lower bound. Mimicking the constructions of the hypercubic subclasses in Assouad's theorem often provides us with important clues as to the construction of a minimax-optimal estimate.

**Birge's modified histogram estimate** is defined by

$$f_n(x) = \frac{\mu_n(A_i)}{\lambda(A_i)} \quad , x \in A_i = [x_{i-1}, x_i) \; , \; i = 1,2,\ldots,m \quad ,$$

where $m$ is a parameter,

$$x_0 = 0 \; ,$$

$$x_i - x_{i-1} = \lambda(A_i) = \frac{\epsilon(1+\epsilon)^{i-1}}{(1+\epsilon)^m - 1} \; ,$$

and $\epsilon > 0$ is another parameter. Observe that $x_m = 1$ and that $\sum_{i=1}^{m} \lambda(A_i) = 1$. Thus, the histogram has geometrically increasing interval sizes. A typical estimate is shown in figure 8.6.

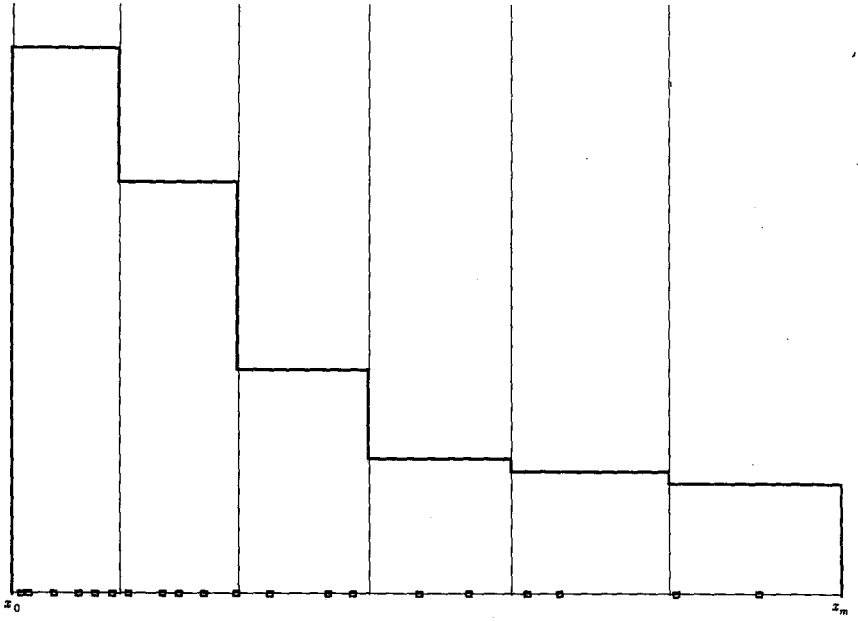


**Figure 8.6.**
Modified histogram estimate with 20 data points and $m = 6$.

**Theorem 8.6. (Birge, 1984)**

Birge's modified histogram estimate on [0,1] with

$$m = \left\lceil \left(4nS^2\right)^{\frac{1}{3}} \right\rceil ,$$

$$\epsilon = e^{\frac{S}{m}} - 1 ,$$

satisfies

$$\sup_{f \in \mathbf{F}} E\left(\int |f_n - f|\right) \leq \frac{3}{2}\left(\frac{2S}{n}\right)^{\frac{1}{3}} + \frac{\sqrt{e}}{8}\left(\frac{2S}{n}\right)^{\frac{2}{3}} ,$$

for all $n \geq 2S$.

If we recall that the lower bound for any estimator over the class $\mathbf{F}$ is about $(1/8)(4S/n)^{1/3}$, we conclude that Birge's estimate is minimax-optimal. Of course, the estimator is only useful when $B$ (and thus $S$) is known beforehand. One may wonder what happens when $S$ is poorly estimated. For fixed $S$, $m$ grows as $n^{1/3}$, and it can be shown that the estimate is consistent for all densities on [0,1]. The rate of convergence can be poor for some monotone densities and many non-monotone densities, because (not surprisingly) the estimator does not take the smoothness of $f$ into account. For monotone bounded densities, one can wonder whether there exists an estimate $f_n$ for which

$$\limsup_{n \to \infty} \sup_{B} \left(\frac{n}{\log(1+B)}\right)^{\frac{1}{3}} \sup_{f \in \mathbf{F}=M_B} E\left(\int |f_n - f|\right) \leq c < \infty$$

for some constant $c$? Note that the estimate itself cannot possibly use the knowledge of $B$. In other words, it has to be adaptive, yet it should share the minimax optimality with Birge's modified histogram estimate. This could be done if we could estimate $B$ satisfactorily from the data. For the details, we refer to Birge (1984).

**Proof of Theorem 8.6.**

Let us introduce the notation $g_i = \int_{A_i} f / \lambda(A_i)$. Then

$$\int |f_n - f| = \sum_{i=1}^{m} \int_{A_i} |f_n - f|$$

$$\leq \sum_{i=1}^{m} \int_{A_i} |f_n - g_i| + \sum_{i=1}^{m} \int_{A_i} |f - g_i|$$

$$\stackrel{\Delta}{=} \text{VARIATION} + \text{BIAS} .$$

Without yet using the exact value of $m$ (because we want to show how the value is obtained), we have

$$\text{BIAS} \leq \sum_{i=1}^{m} \frac{1}{2}\lambda(A_i)(f(x_{i-1})-f(x_i))$$

$$= \frac{1}{2}\left( f(0)\lambda(A_1) + \sum_{i=2}^{m}(1+\epsilon)f(x_{i-1})\lambda(A_{i-1}) - \sum_{i=1}^{m} f(x_i)\lambda(A_i) \right)$$

$$\leq \frac{1}{2}\left( B\lambda(A_1) + \epsilon \sum_{i=2}^{m} f(x_{i-1})\lambda(A_{i-1}) \right)$$

$$\leq \frac{\epsilon}{2}\left( 1+\frac{B}{(1+\epsilon)^m - 1} \right),$$

where only the first inequality requires explanation. What we used there was the fact that if $f$ is a $\downarrow$ density on $[a,b]$ with average $\bar{f}$, then

$$\int_a^b |f-\bar{f}| \leq \frac{1}{2}(b-a)(f(a)-f(b))$$

as can be seen from figure 8.7.

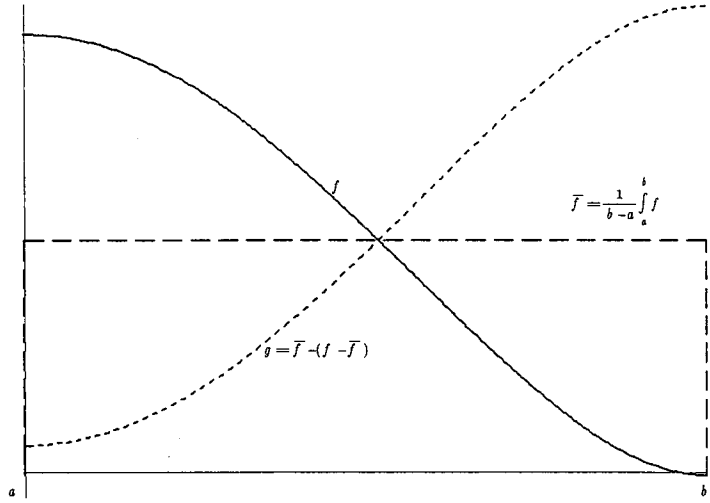


Figure 8.7.

In this figure, $g = \bar{f} - (f - \bar{f})$ is nothing but the mirror image of $f$ about the average $\bar{f}$. Thus,

$$\int_a^b |f - \bar{f}| = \frac{1}{2}\int_a^b |f - g| \leq \frac{1}{2}(b-a)\frac{2(f(a) - \bar{f}) + 2(\bar{f} - f(b))}{2}$$

$$= \frac{1}{2}(b-a)(f(a) - f(b)).$$

The last inequality follows easily from a standard monotonicity argument as illustrated in figure 8.8.

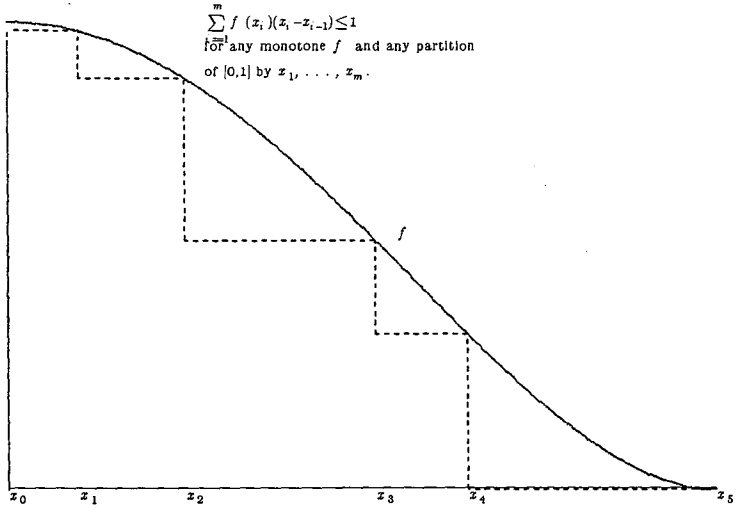


**Figure 8.8.**

To bound the VARIATION term, we let $Z_i$ be a binomial random variable with parameters $n$ and $p_i = \int_{A_i} f$, and proceed as follows:

$$
\begin{aligned}
E(\text{VARIATION}) &= \sum_{i=1}^{m} \frac{1}{n\,\lambda(A_i)} \lambda(A_i) E(\,|\,Z_i - E(Z_i)\,|\,) \\
&\le \frac{1}{n} \sum_{i=1}^{m} \sqrt{E((Z_i - E(Z_i))^2)} \quad \text{(Cauchy-Schwarz)} \\
&= \frac{1}{n} \sum_{i=1}^{m} \sqrt{np_i(1-p_i)} \\
&= \frac{m}{\sqrt{n}} \frac{1}{m} \sum_{i=1}^{m} \sqrt{p_i(1-p_i)} \\
&\le \frac{m}{\sqrt{n}} \sqrt{\frac{1}{m} \sum_{i=1}^{m} p_i (1 - \frac{1}{m} \sum_{i=1}^{m} p_i)} \quad \text{(Jensen)} \\
&= \sqrt{\frac{m-1}{n}}\,.
\end{aligned}
$$

Combining the bounds on the bias and the variation, we have

$$E\left(\int |f_n - f|\right) \le \frac{\epsilon}{2}\left(1 + \frac{B}{(1+\epsilon)^m - 1}\right) + \sqrt{\frac{m-1}{n}}$$

$$= \frac{\epsilon}{2}\left(1 + \frac{B}{e^S - 1}\right) + \sqrt{\frac{m-1}{n}} \quad (\epsilon = e^{\frac{S}{m}} - 1)$$

$$= \epsilon + \sqrt{\frac{m-1}{n}} \quad (\text{definition of } S = \log(1+B))$$

$$\le e^{\frac{S}{x}} - 1 + \sqrt{\frac{x}{n}} \quad (\text{any } m-1 < x \le m).$$

The derivative of the last expression with respect to $x$ is zero for the solution of

$$x^{\frac{3}{2}} = 2S\sqrt{n}\, e^{\frac{S}{x}}.$$

Roughly speaking, this solution increases with $n$, so that $e^{\frac{S}{x}}$ approaches 1 as $n \to \infty$. Therefore, the choice

$$x = \left(4nS^2\right)^{\frac{1}{3}}, \quad m = \lceil x \rceil$$

is reasonable. Plugging this back into the upper bound for the expected $L_1$ error, we obtain

$$E\left(\int |f_n - f|\right) \le \sqrt{\frac{x}{n}} + \frac{S}{x} + \frac{S^2}{2x^2} e^{\frac{S}{x}} \quad (\text{use } e^u - 1 \le u + \frac{u^2}{2} e^u)$$

$$\le \left(\frac{2S}{n}\right)^{\frac{1}{3}} \left(1 + \frac{1}{2}\right) + \frac{1}{2}\left(\frac{2S}{n}\right)^{\frac{2}{3}} \frac{1}{4} e^{\frac{1}{2}\left(\frac{2S}{n}\right)^{\frac{1}{3}}}$$

$$\le \frac{3}{2}\left(\frac{2S}{n}\right)^{\frac{1}{3}} + \frac{\sqrt{e}}{8}\left(\frac{2S}{n}\right)^{\frac{2}{3}} \quad (\text{if } 2S \le n). \ \blacksquare$$

## 8.6. EXERCISES.

8.1. This exercise deals with an association inequality needed in the proof of Theorem 8.2. Let $h_1 \ge h_2 \ge \cdots \ge h_n$ be nonnegative numbers, and assume that $\Delta_1, \ldots, \Delta_n$ are positive numbers with the property that for all $1 \le i \le n$,

$$\frac{\Delta_1 + \cdots + \Delta_i}{i} \ge \frac{\Delta_1 + \cdots + \Delta_n}{n}.$$

Thus, $\Delta_i$ has a tendency to decrease in $i$, but may not be monotone itself. Show that

$$\frac{1}{n}\sum_{i=1}^{n} h_i \Delta_i \geq \frac{1}{n}\sum_{i=1}^{n} h_i \ \frac{1}{n}\sum_{i=1}^{n} \Delta_i \ .$$

Hint: use induction on $n$.

8.2. This exercise is about the minimax optimality of Grenander's estimate $f_n$. Let $\mathbf{F}$ be the class of all monotone densities on [0,1], bounded by $B$. Is the estimate asymptotically minimax-optimal, in the sense that there exists a constant $C$ such that

$$\limsup_{n\to\infty} \sup_{\mathbf{F}} n^{\frac{1}{3}} E\left(\int |f_n - f|\right) \leq C\, S^{\frac{1}{3}}$$

where we recall that $S = \log(B+1)$?

8.3. Show that for the uniform density on [0,1], Grenander's estimate has an $O(1/\sqrt{n})$ expected $L_1$ error. (The same is true for all finite histogram-shaped densities.)

8.4. Let $f$ be a density on $[0,\infty)$. The ML histogram based upon the data is a histogram with breakpoints at the data points and at the origin such that on $(X_{(i-1)}, X_{(i)}]$, its value is $1/(n(X_{(i)} - X_{(i-1)}))$, where $X_{(i)}$ is the $i$-th order statistic. Its distribution function is obtained by joining the upper vertices of the empirical distribution function by straight lines. See figure 8.9 below. Show that for **all** $f$,

$$\liminf_{n\to\infty} E\left(\int |f_n - f|\right) > 0 \ .$$

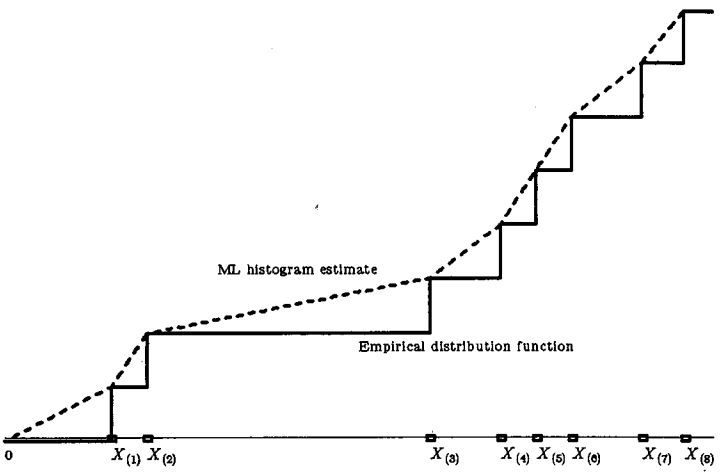


**Figure 8.9.**

The empirical distribution function and the distribution function of the ML histogram estimate are shown.

# Chapter Nine
# RELATIVE STABILITY

## 9.1. DEFINITION AND MOTIVATION.

A density estimate $f_n$ is **relatively stable** when

$$\frac{J_n}{E(J_n)} \to 1 \text{ in probability as } n \to \infty ,$$

where $J_n = \int |f_n - f|$. It is **strongly relatively stable** when the convergence is in the almost sure sense.

The notion of relative stability is important in comparative studies of density estimates. Comparing relatively stable density estimates on the basis of $E(J_n)$ is fair since the actual error $J_n$ is with high probability close to its mean. The situation is more complicated for example when $J_n / E(J_n)$ tends to a nondegenerate limit law; the conservative elements among us could be tempted to choose a density estimate with a larger asymptotic mean but a smaller asymptotic variance. Dilemmas of this sort do not occur for relatively stable density estimates.

Another important point is that simulations of the performance ($L_1$ error) of relatively stable density estimates are very cheap since $J_n$ (the computed performance) is with high probability close to $E(J_n)$. In other words, it is not necessary to average over several simulation runs. As we will see below, $J_n$ is already an average of sorts because of the global integral in its definition.

The literature on minimax lower and upper bounds for the $L_1$ error deals almost exclusively with $E(J_n)$, and not with other quantities such as the $p$-th quantiles of $J_n$. In view of the relative stability of most nonparametric estimates, it is less important to develop minimax theories based upon quantities other than $E(J_n)$, except in special circumstances. One such situation is when the classes of densities considered in the minimax theory are very small ("parametric"), so that specially designed estimates ("parametric estimates") are better suited.

Most parametric density estimates are not relatively stable. Take for example the class of densities $f = pg + (1-p)h$ where $g, h$ are known disjoint densities

($\int gh = 0$), and $p$ is the unknown mixture parameter. If $p$ is estimated from the data by the obvious frequency estimate $p_n$, and $f_n = p_n g + (1-p_n)h$, then $J_n = 2 \mid p - p_n \mid$, and thus, by the central limit theorem,

$$\frac{J_n}{2\sqrt{p\,(1-p\,)/n}} \to \mid N \mid \text{ in distribution}$$

where $N$ is a normal random variable. It is clear that $J_n / E(J_n) \to \mid N \mid / E(\mid N \mid)$ in distribution as $n \to \infty$. Therefore, the estimate is not relatively stable for any density in the given class.

In contrast, popular nonparametric density estimates such as the kernel and histogram estimates are relatively stable for all densities. This is due to the local nature of these estimates: densities are estimated locally by considering a limited number of close data points. Locally, the error's standard deviation can be of the same order of magnitude as the error's mean. Yet, because the $L_1$ criterion sums a lot of many "nearly independent" local errors, the variation in the local errors averages out, rendering the estimates relatively stable. Thus, if we had picked a local criterion such as $\mid f_n - f \mid / E(\mid f_n - f \mid)$ to define relative stability, then relative stability would effectively force the bias term to dominate the variational term ,i.e. $E(\mid f_n - f \mid) \sim \mid E(f_n) - f \mid$. For nonparametric density estimates, one can usually achieve this by taking the smoothing parameter large enough. Yet, this is a suboptimal strategy because the smallest asymptotic errors are obtained by balancing the bias and variation terms. Thus, "local relative stability" and "locally optimal rate of convergence" are conflicting notions.

We will show that for the kernel estimate, consistency implies relative stability. It suffices to note that everything that follows remains valid for the histogram estimate as well. We consider smoothing factors $h$ that are functions of $n$ only. Relative stability in the $L_2$ sense (replace $J_n$ by $\int (f_n - f)^2$ in the definition of relative stability) has been established by Hall (1982) under some regularity conditions on $h$, $f$ and $K$, when $d = 1$. Later, Hall (1984) refined this result and obtained the asymptotic law of $\int (f_n - f)^2$ when $f$ has two uniformly continuous derivatives on $R^d$. Unfortunately, for a variety of reasons, $L_1$ relative stability cannot be obtained from Hall's results.

## 9.2. MAIN RESULTS.

We consider only **boxed kernels K**, i.e. kernels $K$ which vanish outside $[-1,1]^d$, and are in absolute value bounded by some finite value $K_{\max}$. Recall that $\int K = 1$, but that $K$ does not have to be nonnegative. The main result from which most other results will be derived is given in Theorem 1:

**Theorem 9.1.**

Consider a kernel estimate with boxed kernel $K$. There exists a universal function $\Psi:(0,\infty)\to(0,\infty)$ not depending upon $f$, $K$ or $d$ such that for all $\epsilon > 3\int |K| / \sqrt{n}$,

$$\sup_{h>0,f} P(|J_n - E(J_n)| > \epsilon)$$
$$\leq 2e^{-\frac{n\epsilon^2}{18\int^2 |K| + 6\epsilon \int |K|}} + 2^{d+1} \inf_{c>0} e^{2^d \Psi(cK_{\max}) - \frac{c\sqrt{n}\,\epsilon}{2^d 3}}.$$

We emphasize that Theorem 1 is valid for **all densities** on $R^d$. Also, the inequality is uniform over all $h$ and all densities $f$. It is true that we pay a price for the uniformity; for particular cases, better inequalities are obtainable. Yet, it is the uniformity that can allow one to establish the relative stability of automatic kernel estimates (see Devroye, 1986). For the uniform kernel in the unit hypercube, taking $c=1$, we obtain the following upper bound, valid for all $\epsilon > 3/\sqrt{n}$:

$$2e^{-\frac{n\epsilon^2}{18+6\epsilon}} + 2^{d+1} e^{2^d \Psi(1)} e^{-\frac{\sqrt{n}\,\epsilon}{2^d 3}}.$$

The long proof of Theorem 9.1 can be avoided if one is willing to accept a weaker inequality. The bound $4\int^2 |K| / (n\epsilon^2)$, obtained in exercise 9.5, is valid for all $f$, $K$, $h$, $n$ and $\epsilon$.

**Corollary 9.1.**

Taking $\epsilon = u/\sqrt{n}$ for some constant $u$, and choosing $c=1$, we obtain for all densities $f$, for all sequences of smoothing factors $h=h_n$, and for $u > 3\int |K|$,

$$\limsup_{n\to\infty} \sup_{h>0,f} P(\sqrt{n}\,|J_n - E(J_n)| > u)$$
$$\leq 2e^{-\frac{u^2}{18\int^2 |K|}} + 2^{d+1} e^{2^d \Psi(K_{\max})} e^{-u/(3\,2^d)}.$$

**Corollary 9.2.**

The simple formula $E(|X|)=\int_0^\infty P(|X|>t)\,dt$ can be used to show that

$$E(|J_n - E(J_n)|) \leq \frac{C}{\sqrt{n}}$$

where we can take

$$C = 3\int |K| + \int_0^\infty (2\, e^{-\frac{t^2}{18\int^2 |K| + 6\frac{\sqrt{t}}{\sqrt{n}}\int |K|}} + 2^{d+1}\, e^{2^d \Psi(cK_{\max})}\, e^{-\frac{ct}{2^d 3}})\,dt\ .$$

The constant $C$ is minimized with respect to $K$ when $K$ is the uniform density on $[-1,1]^d$, so that $K_{\max}=2^{-d}, \int |K| = 1$. As $n \to \infty$, $C$ tends to a constant depending upon $c$ and the kernel only. Thus, $|J_n - E(J_n)|$ decreases to zero roughly as $1/\sqrt{n}$, which is much faster than the rate with which $E(J_n)$ tends to zero, i.e. $n^{-2/5}$ or slower for any nonnegative kernel $K$ and $d=1$.

**Corollary 9.3.**

If we take $\epsilon = \log(n)/\sqrt{n}$, and choose $c$ very large in Theorem 9.1, then it is easily seen (apply the Borel-Cantelli theorem) that for all $f$ and all sequences of smoothing factors $h = h_n$,

$$\sqrt{n}\ |J_n - E(J_n)| = o(\log(n)) \text{ almost surely as } n \to \infty\ .$$

**Corollary 9.4.**

For symmetric nonnegative boxed kernels, and all $f$, we have

$$\liminf_{n\to\infty} \inf_h n^{\frac{2}{5}} E(J_n) \geq \gamma > 0$$

for some universal constant $\gamma$ which is at least equal to 0.8 (Devroye and Gyorfi, 1985, p.79). Combining this with the previous remarks yields strong relative stability. In fact, we have more information: regardless of how $h = h_n$ is chosen as a function of $n$, we have

$$\frac{|J_n - E(J_n)|}{E(J_n)} = o(n^{-\frac{1}{10}} \log(n)) \text{ almost surely as } n \to \infty\ .$$

**Corollary 9.5.**

For any boxed kernel and any density $f$ on $R^1$ and any sequence of smoothing factors $h \to 0$, the kernel estimate is relatively stable. We used the fact that for all boxed kernels $K$, $h \to 0$ implies $\sqrt{n}\, E(J_n) \to \infty$ as $n \to \infty$ (see exercise 9.2; or Devroye and Gyorfi (1985, p. 136)). One question not answered is whether the condition $h \to 0$ can be dropped altogether. This is dealt with in exercise 9.1.

## 9.3. A MOMENT INEQUALITY FOR THE POISSON DISTRIBUTION.

We will prove Theorem 9.1 via a Poissonization argument. At a crucial junction, it will be necessary to bound $E(N^p)$ from above where $p$ is a large integer, and $N$ is a Poisson $(n)$ random variable. The bound should be explicit in $n$ and $p$, and should not increase rapidly with $p$.

**Lemma 9.1.**

Let $N$ be a Poisson $(n)$ random variable, and let $p \geq 1$ be a constant. Then

$$E(N^p) \leq \left(\frac{p}{e}\right)^p \frac{e^{p/\log(1+p/n)}}{\log^p\left(1+\frac{p/n}{\log(1+p/n)}\right)}$$

and

$$E(N^p) \leq n^p \left(\frac{p}{\log(1+p)}\right)^p .$$

Furthermore,

$$E((N-n)^p) \leq \left(\frac{n}{e}\right)^{\frac{p}{2}} \left(\frac{p}{\log(1+\sqrt{p})}\right)^p .$$

**Proof of Lemma 9.1.**

For any nonnegative integer $i$ and numbers $\delta > 1, p > 0$, we have

$$i^p \leq \delta^i \left(\frac{p}{\log(\delta)}\right)^p \delta^{-p/\log(\delta)} = \delta^i \left(\frac{p}{e\log(\delta)}\right)^p .$$

This follows after observing that the function $x^p \delta^{-x}$ is maximal at the value $x = p/\log(\delta)$. Thus,

$$E(N^p) = \sum_{i=0}^{\infty} i^p \frac{e^{-n} n^i}{i!}$$

$$\leq \left(\frac{p}{e \log(\delta)}\right)^p e^{n(\delta-1)} \sum_{i=0}^{\infty} \frac{e^{-n\delta}(n\delta)^i}{i!} = \left(\frac{p}{e \log(\delta)}\right)^p e^{n(\delta-1)}.$$

We still have the freedom to choose $\delta$. The expression is minimal when $\delta$ is the solution of $\delta \log(\delta) = p/n$. Since we want $\delta > 1$, we should choose $\delta$ for large values of $p/n$ approximatively as follows:

$$\delta = 1 + \frac{p/n}{\log(1+p/n)}.$$

This gives the first inequality. Since $\delta \log(\delta) \approx \delta - 1$ for $\delta$ near 1, we should pick $\delta = 1 + \frac{p}{n}$ when $p/n$ is small. The latter choice gives the inequality

$$E(N^p) \leq e^p \left(\frac{p}{e \log(1+\frac{p}{n})}\right)^p = (np)^p (n \log(1+\frac{p}{n}))^{-p}$$

$$\leq (np)^p / \log^p(1+p).$$

Since this yields an adequate inequality, the former choice of $\delta$ will not be explored any further. To show the last inequality, we begin with $N^p \leq 2^{p-1}(n^p + (N-n)^p)$, and observe that for all $\delta > 1$,

$$E((N-n)^p) \leq e^{n(\delta-1-\log(\delta))} \left(\frac{p}{e \log(\delta)}\right)^p$$

(argue as before: $(x-n)^p \delta^{-x}$ is maximal for $x = n + p/\log(\delta)$). The bound is minimal (in $\delta$) when $\delta$ is the solution of $(\delta-1)\log(\delta) = p/n$. For small $p/n$, the solution is close to 1. Since $\log(\delta) \approx \delta - 1$, we can take in first approximation $\delta = 1 + \sqrt{p/n}$. Resubstitution of this value and using the fact that $\delta - 1 - \log(\delta) \leq (\delta-1)^2/2$ yields

$$E((N-n)^p) \leq e^{p/2} \left(\frac{p}{e \log(\delta)}\right)^p$$

$$\leq (ne)^{p/2} \left(\frac{p}{e\sqrt{n}\, \log(1+\sqrt{p/n})}\right)^p$$

$$\leq \left(\frac{n}{e}\right)^{p/2} \left(\frac{p}{\log(1+\sqrt{p})}\right)^p. \quad \blacksquare$$

## 9.4. TWO FUNDAMENTAL TOOLS.

One of the problems with a random variable of the form $\int |f_n - E(f_n)|$ is that when the integral is written as a sum of very many integrals of disjoint sets, the terms in this sum are dependent because of two phenomena: first of all, if $K$ has support contained in $[-1,1]^d$, then a particular $X_i$ contributes to $f_n(x)$ for all $x$ within the hypercube $X_i + [-h,h]^d$. Thus, at the very least, the integrals over neighboring sets in the partition are dependent. This dependence will be taken care of by grouping the sets into $2^d$ classes: within each class, the sets are all sufficiently far apart. The second phenomenon is of a multinomial nature: the cardinalities of the sets (in terms of the $X_i$'s) are multinomially distributed. We can either work with the multinomial random variables or use a Poissonization argument. Since the former approach is a bit longer, we will employ Poissonization to make the cardinalities of the sets independent of one another.

Let us first establish our notation. $X_1, X_2, \ldots$ is an infinite sequence of iid random vectors with common density $f$. $N$ is a Poisson $(n)$ random variable, independent of the infinite data sequence. We define

$$f_N = \frac{1}{n} \sum_{i \le N} K_h(x - X_i),$$

$$J_N = \int |f_N - f|.$$

The space $R^d$ is partitioned into sets indexed by $d$-tuples $\alpha = (j_1, \ldots, j_d)$ where all $j_i$'s are integer-valued. The set $A_\alpha$ is defined by

$$A_\alpha = \prod_{i=1}^{d} [j_i 2h, (j_i+1)2h).$$

The $d$-tuples $\alpha$ are partitioned in turn into $2^d$ classes $C_1, \ldots, C_{2^d}$ according to the $2^d$ possible odd-even patterns for the $d$ integer components of $\alpha$. See figure 9.1.

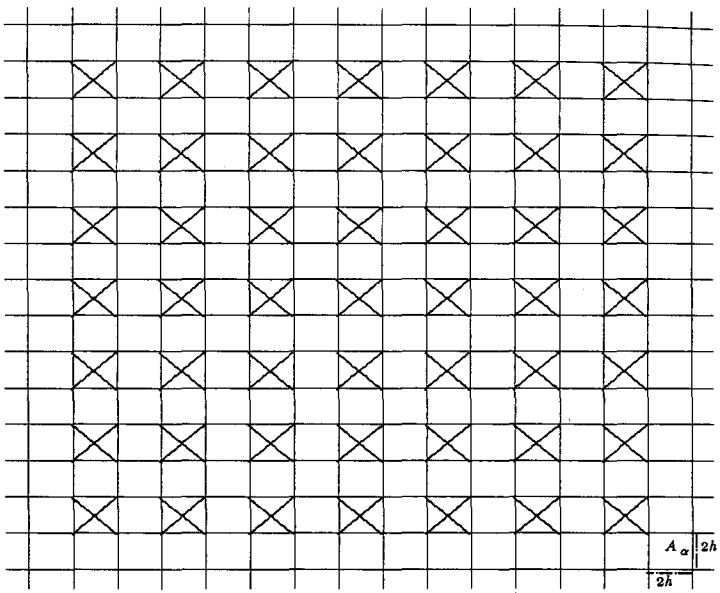


**Figure 9.1.**

Grid with $A_\alpha = [2hj_1, 2h(j_1+1)) \times \cdots \times [2hj_d, 2h(j_d+1))$, where $\alpha = (j_1, \ldots, j_d)$.

One class of sets $A_\alpha$ is marked with crosses.

We define

$$Y_\alpha = \int_{A_\alpha} (|f_N - f| - E(|f_N - f|)).$$

It should be noted that when $\alpha, \beta \in C_i$, and $\alpha \neq \beta$, then $Y_\alpha$ and $Y_\beta$ are independent. Notice that we have partitioned the space into $2^d$ grids, where each grid is a translation of the first grid.

The second tool needed in our proof is an inequality linking $J_n = \int |f_n - f|$ to $J_N = \int |f_N - f|$:

**Lemma 9.2.**

For all $\epsilon > 3\int |K|/\sqrt{n}$,

$$P(|J_n - E(J_n)| > \epsilon) \leq P(|J_N - E(J_N)| > \frac{\epsilon}{3}) + 2e^{-\frac{n\epsilon^2}{18\int^2 |K| + 6\epsilon \int |K|}}.$$

**Proof of Lemma 9.2.**

By the triangle inequality,

$$P(|J_n - E(J_n)| > \epsilon)$$
$$\leq P(|J_N - E(J_N)| > \frac{\epsilon}{3}) + P(|J_N - J_n| > \frac{\epsilon}{3}) + P(|E(J_n) - E(J_N)| > \frac{\epsilon}{3}).$$

Observe that

$$|J_n - J_N| = |\int |f_n - f| - \int |f_N - f||$$
$$\leq \int |f_n - f_N| \leq \frac{|N-n|}{n} \int |K|.$$

Also, $\frac{E(|N-n|)}{n} \int |K| \leq n^{-1/2} \int |K| < \epsilon/3$. Furthermore,

$$P\left( \frac{|N-n|}{n} \int |K| > \frac{\epsilon}{3} \right)$$
$$= P(|N-n| > \delta n) \quad (\delta \stackrel{\Delta}{=} \frac{\epsilon}{3\int |K|})$$
$$\leq E(e^{t|N-n| - t\delta n}) \quad (t > 0) \quad \text{(Jensen's inequality)}$$
$$\leq e^{-t\delta n} E\left( e^{t(N-n)} + e^{t(n-N)} \right)$$
$$= e^{-t\delta n} \left( e^{n(e^t - 1 - t)} + e^{n(t + e^{-t} - 1)} \right)$$
$$\leq 2\, e^{-t\delta n + n(e^t - 1 - t)} \quad (\text{since } e^{-t} + t \leq e^t - t)$$
$$= 2\, e^{n(\delta - (1+\delta)\log(1+\delta))} \quad (t = \log(1+\delta))$$
$$\leq 2\, e^{-\frac{n\delta^2}{2(1+\delta)}}.$$

The last inequality is most easily shown via Taylor's series expansion. ■

We conclude this section by giving two other inequalities:

**Lemma 9.3.**

Let $X$ be any random variable with finite mean, and let $a$ be an arbitrary real number. Then

$$| \, | X-a \, | - E( | X-a \, | ) | \leq | X-E(X)) | + E( | X-E(X) | ) .$$

**Proof of Lemma 9.3.**

Observe that

$$\begin{aligned} & | X-a \, | - | X-E(X) | \\ \leq & | a-E(X) | \leq E( | X-a \, | ) , \end{aligned}$$

and that

$$\begin{aligned} & E( | X-a \, | ) - E( | X-E(X) | ) \leq E( | a-E(X) | ) \\ = & | a-E(X) | \leq | a-X \, | + | X-E(X) | \, . \blacksquare \end{aligned}$$

**Lemma 9.4. Whittle's inequality.**

Let $X_1, \ldots, X_n$ be iid zero mean random variables with finite $p$-th moment, where $p \geq 2$ is a real number. Then

$$E( | \frac{1}{n} \sum_{i=1}^{n} X_i \, |^p ) \leq \gamma_p \, n^{-\frac{p}{2}}$$

where

$$\gamma_p = \frac{1}{\sqrt{\pi}} \, 8^{\frac{p}{2}} \Gamma(\frac{p+1}{2}) \, .$$

( The general form of the inequality is due to Marcinkiewicz and Zygmund (1937), and the constant $\gamma_p$ was computed by Whittle (1960).)

**Proof of Lemma 9.4.**

See Whittle (1960). ■

## 9.5. PROOF OF THEOREM 9.1.

We begin with

$$J_N - E(J_N) = \sum_{i=1}^{2^d} \sum_{\alpha \in C_i} Y_\alpha .$$

Thus, for all $\epsilon > 0$,

$$P(|J_N - E(J_N)| > \epsilon) \leq \sum_{i=1}^{2^d} P\left( \left| \sum_{\alpha \in C_i} Y_\alpha \right| > \frac{\epsilon}{2^d} \right)$$

$$\leq \sum_{i=1}^{2^d} e^{-t\frac{\epsilon}{2^d}} \left( \prod_{\alpha \in C_i} E(e^{tY_\alpha}) + \prod_{\alpha \in C_i} E(e^{-tY_\alpha}) \right) \quad (\text{all } t > 0) .$$

Here we used Jensen's inequality and the independence of the $Y_\alpha$'s within the same class $C_i$. This is the place in which the shifted grid argument is fully exploited.

Everything now boils down to an investigation of $E(e^{tY_\alpha})$ for all $t$. From Taylor's series with remainder term, we have

$$E(e^{tY_\alpha}) \leq E\left( 1 + \sum_{j=2}^{\infty} \frac{|t|^j |Y_\alpha|^j}{j!} \right) .$$

Observe that

$$|Y_\alpha| \leq \int_{A_\alpha} \Big| |f_N - f| - E(|f_N - f|) \Big|$$

$$\leq \int_{A_\alpha} \Big( |f_N - E(f_N)| + E(|f_N - E(f_N)|) \Big)$$

(Lemma 9.3)

$$\stackrel{\Delta}{=} Z_\alpha .$$

The last inequality follows by repeated applications of the triangle inequality. Combining this with the fact that $E(Y_\alpha) = 0$ shows that

$$E(e^{tY_\alpha}) \leq E\left( 1 + \sum_{j=2}^{\infty} \frac{|t|^j Z_\alpha^j}{j!} \right) .$$

Note that

$$Z_\alpha^j = \left( \int_{A_\alpha} \Big( |f_N - E(f_N)| + E(|f_N - E(f_N)|) \Big) \right)^j$$

$$\leq \lambda^{j-1}(A_\alpha) \int_{A_\alpha} \Big( |f_N - E(f_N)| + E(|f_N - E(f_N)|) \Big)^j$$

(Holder's inequality)

$$\leq (2h)^{(j-1)d} 2^{j-1} \int_{A_\alpha} \Big( |f_N - E(f_N)|^j + E^j(|f_N - E(f_N)|) \Big) .$$

Taking expected values and applying Jensen's inequality shows that

$$E(Z_\alpha^j) \leq (2h)^{(j-1)d} 2^j \int_{A_\alpha} E\Big( |f_N - E(f_N)|^j \Big) .$$

Let $J$ be the set of indices $i$ for which $n < i \leq N$ or $N < i \leq n$. Observe that

$$E\Big( |f_N - E(f_N)|^j \Big)$$

$$= E\Big( |f_N - E(f_n)|^j \Big) \quad (Ef_N = E(N)E(K_h) = nE(K_h))$$

$$\leq 2^{j-1} \left( E\left( \left| \frac{1}{n} \sum_{i=1}^{n} (K_h(x - X_i) - E(K_h(x - X_i))) \right|^j \right) + E\left( \left| \frac{1}{n} \sum_{i \in J} K_h(x - X_i) \right|^j \right) \right) .$$

The first expected value in the last expression is not larger than

$$\gamma_j n^{-\frac{j}{2}} E\Big( |K_h(x - X_1) - E(K_h(x - X_1))|^j \Big) \quad \text{(Lemma 9.4)}$$

$$\leq \gamma_j n^{-\frac{j}{2}} 2^j E\Big( |K_h(x - X_1)|^j \Big)$$

$$= 2^j \gamma_j n^{-\frac{j}{2}} |K_h|^j * f .$$

The second expected value is not larger than

$$n^{-j} E\left( |N - n|^j \sum_{i=1}^{|N-n|} |K_h(x - X_i)|^j \right) \quad \text{(Jensen)}$$

$$= n^{-j} E\Big( |K_h(x - X_1)|^j \Big) E(|N - n|^j)$$

$$= n^{-j} E(|N - n|^j) \, |K_h|^j * f$$

$$\leq n^{-j} \xi_j n^{\frac{j}{2}} |K_h|^j * f$$

$$\left( \xi_j \stackrel{\Delta}{=} \left( \frac{j}{\sqrt{e} \log(1 + \sqrt{j})} \right)^j \right) \text{(Lemma 9.1)} .$$

Thus,

$$E(Z_\alpha^j) \le (2h)^{(j-1)d} 2^j \int_{A_\alpha} 2^{j-1}\left(2^j \gamma_j + \xi_j\right) n^{-\frac{j}{2}} \mid K_h \mid^j * f$$

$$\stackrel{\Delta}{=} \rho_j \, h^{(j-1)d} \, n^{-\frac{j}{2}} \int_{A_\alpha} \mid K_h \mid^j * f \ .$$

Therefore,

$$E(e^{tY_\alpha}) \le 1 + \sum_{j=2}^{\infty} \left(\frac{t^2}{n}\right)^{\frac{j}{2}} \frac{\rho_j}{j!} \, h^{(j-1)d} \int_{A_\alpha} \mid K_h \mid^j * f$$

$$\stackrel{\Delta}{=} 1 + G(\alpha,t) \le e^{G(\alpha,t)} \ .$$

Note that $G(\alpha,t)$ is an even function of $t$, so that $Ee^{-tY_\alpha}$ can be bounded by the same expression. Thus,

$$P(\mid J_N - E(J_N) \mid \ > \epsilon) \le \sum_{i=1}^{2^d} 2e^{-\mid t \mid \epsilon/2^d} \prod_{\alpha \in C_i} e^{G(\alpha,t)}$$

$$= \sum_{i=1}^{2^d} 2e^{-\mid t \mid \epsilon/2^d} \, e^{\sum_{\alpha \in C_i} G(\alpha,t)}$$

$$\le 2^{d+1} e^{-\mid t \mid \epsilon/2^d} \, e^{\sum_{j=2}^{\infty} \left(\frac{t^2}{n}\right)^{\frac{j}{2}} \frac{\rho_j}{j!} 2^d (K_{\max})^j}$$

where we used the fact that $\sum_{\alpha \in C_i} \int_{A_\alpha} \mid K_h \mid^j * f \le \int \mid K_h \mid^j * f = \int \mid K \mid^j h^{-(j-1)d} \le 2^d K_{\max}^j h^{-(j-1)d}$. Putting $t^2 = nc^2$, we obtain

$$P(\mid J_N - E(J_N) \mid \ > \epsilon) \le 2^{d+1} e^{-c\sqrt{n}\,\epsilon/2^d}$$

$$\stackrel{\Delta}{=} 2^{d+1} e^{-c\sqrt{n}\,\epsilon/2^d} \, e^{2^d \Psi(cK_{\max})} \ .$$

We need only show that for all positive constants $u$, $\Psi(u) < \infty$. By the definitions of $\xi_j$ and $\gamma_j$, this is equivalent to showing that for all $u > 0$,

$$\sum_{j=2}^{\infty} \left(\frac{j}{\log(1+\sqrt{j})}\right)^j \frac{u^j}{j!} < \infty \ ,$$

and

$$\sum_{j=2}^{\infty} \Gamma\left(\frac{j+1}{2}\right) \frac{u^j}{j!} < \infty \ .$$

This can be verified quite easily using Stirling's approximation for the factorial. This concludes the proof of Theorem 9.1. ∎

## 9.6. EXERCISES.

9.1. Show that there exist densities $f$ and kernels $K$ such that the kernel estimate with constant smoothing factor $h$ is not relatively stable.

9.2. Show that for all boxed kernels $K$ and all sequences of smoothing factors $h \downarrow 0$, the kernel estimate $f_n$ cannot converge at the rate $1/\sqrt{n}$, i.e,

$$\lim_{n \to \infty} \sqrt{n} \; E\left(\int |f_n - f|\right) = \infty .$$

9.3. This is an introduction to the next exercise. Assume that the $L_1$ distance between two densities $f$ and $g$ is estimated by importance sampling in the following manner: generate $Y_1, \ldots, Y_m$, iid random variables with density $f$. Compute the estimate

$$T_m = \frac{1}{m} \sum_{i=1}^{m} 2 \left| 1 - \frac{g(Y_i)}{f(Y_i)} \right|_+ .$$

Show that $E(T_m) = \int |f - g|$, and that

$$Var(T_m) \leq \frac{2 \int |f - g|}{m} .$$

Furthermore, $P(|T_m - \int |f - g|| > \epsilon)$ is bounded by the minimum of

$$2e^{-\frac{m\epsilon^2}{2}}$$

and

$$2e^{-\frac{m\epsilon^2}{\frac{4}{m}\int |f-g| + 4\epsilon}} .$$

Note: these inequalities provide crude confidence intervals for the $L_1$ error. Compute a safe value for $m$ such that with probability at least 95% $T_m$ is within 5% of its mean, when $\int |f - g|$ takes the values 0.1,0.05,0.02,0.01 (this is the range of values usually of interest in density estimation).

9.4. Give a distribution-free confidence interval for $T_m$, when it is used to estimate $E\left(\int |f_n - f|\right)$ where $f$ is known, $f_n$ is the kernel estimate with boxed kernel $K$ and arbitrary $h$. The estimate $T_m$ is, as in exercise 9.3, based upon a sample of iid random variables drawn from $f_n$. In other words, give a useful distribution-free upper bound for

$$P\left(\left|T_m - E\left(\int |f_n - f|\right)\right| > \epsilon\right) .$$

The bound should depend upon $\epsilon, n, K, m$ and $d$ only. Show also that

$$E\left(\left(T_m - E\left(\int |f_n - f|\right)\right)^2\right) \leq \frac{4}{m} + \frac{C}{n}$$

for some constant $C$ depending upon $K$ and $d$ only.

9.5. Show that for the kernel estimate with arbitrary kernel $K$, we have

$$Var\left(\int |f_n - f|\right) \leq \frac{4\int^2 |K|}{n}.$$

Show that for the histogram estimate and for the kernel estimate with nonnegative kernel, we have

$$Var\left(\int |f_n - f|\right) \leq \frac{4}{n}.$$

Note that this can be used to determine the constant $C$ in exercise 9.4, and that weak relative stability follows from this result whenever $\lim_{n\to\infty} \sqrt{n}\, E\left(\int |f_n - f|\right) = \infty$. Hint: try using the Efron-Stein inequality (Efron and Stein, 1981), which states that

$$Var\left(S(X_1, \ldots, X_n)\right) \leq E\left\{\sum_{i=1}^{n+1} (S_i - \bar{S})^2\right\}.$$

where $X_1, \ldots, X_n, X_{n+1}$ are iid random vectors, $S(x_1, \ldots, x_n)$ is a symmetric function of its arguments,

$$S_i = S(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{n+1}), \quad i = 1, \ldots, n+1,$$

and $\bar{S} = (n+1)^{-1} \sum_{i=1}^{n+1} S_i$.

# REFERENCES


S. Abou-Jaoude, "La convergence L1 et L∞ de certains estimateurs d'une densite de probabilite," These de Doctorat d'Etat, University Paris VI, France, 1977.

R. A. Adams, *Sobolev Spaces,* Academic Press, New York, 1975.

P. Assouad, "Deux remarques sur l'estimation," *Comptes Rendus de l'Academie des Sciences de Paris*, vol. 296, pp. 1021-1024, 1983.

M. S. Bartlett, "Statistical estimation of density functions," *Sankhya Series A*, vol. 25, pp. 245-254, 1963.

A. P. Basu, "Estimates of reliability for some distribution useful in life testing," *Technometrics*, vol. 6, pp. 215-219, 1964.

G. Bennett, "Probability inequalities for the sum of independent random variables," *Journal of the American Statistical Association*, vol. 57, pp. 33-45, 1962.

R. Beran, "Minimum Hellinger distance estimates for parametric models," *Annals of Statistics*, vol. 5, pp. 445-463, 1977.

R. Beran, "Efficient robust estimates in parametric models," *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 55, pp. 91-108, 1981.

P. J. Bickel, "Another look at robustness: a review of reviews and some new developments," *Scandinavian Journal of Statistics*, vol. 3, pp. 145-168, 1976.

L. Birge, "Estimating a density under order restrictions: non-asymptotic minimax risk," Technical Report, UER de Sciences Economiques, Universite Paris X, Nanterre, France, 1983.

L. Birge, "On the risk of histograms for estimating decreasing densities," Technical Report, UER de Sciences Economiques, Universite Paris X, Nanterre, France, 1984.

L. Birge, "Non-asymptotic minimax risk for Hellinger balls," *Probability and Mathematical Statistics*, vol. 5, pp. 21-29, 1985.

L. Birge, "On estimating a density using Hellinger distance and some other strange facts," *Probability Theory and Related Fields*, vol. 71, pp. 271-291, 1986.

J. Bretagnolle and C. Huber, "Lois empiriques et distance de Prokhorov," in *Seminaire de Probabilites XII*, vol. 649, pp. 332-341, Springer-Verlag, New York, 1978.

J. Bretagnolle and C. Huber, "Estimation des densites: risque minimax," *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 47, pp. 119-137, 1979.

P. Burman, "A data dependent approach to density estimation," *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 69, pp. 609-628, 1985.

P. L. Butzer and R. J. Nessel, *Fourier Analysis and Approximation*, Birkhauser Verlag, Basel and Stuttgart, 1971.

T. Cacoullos, "Estimation of a multivariate density," *Annals of the Institute of Statistical Mathematics*, vol. 18, pp. 178-189, 1966.

F. Carlson, "Une inegalite," *Arkiv foer Matematik, Astromi och Fysik*, vol. 25B, pp. 1-15, 1934.

G. F. Clements, "Entropies of several sets of functions," *Pacific Journal of Mathematics*, vol. 13, pp. 1085-1097, 1963.

T. M. Cover, "Learning in pattern recognition," in *Methodologies of Pattern Recognition*, ed. S. Watanabe, pp. 111-132, Academic Press, New York, N.Y., 1969.

T. M. Cover and T. J. Wagner, "Topics in statistical pattern recognition," *Communication and Cybernetics*, vol. 10, pp. 15-46, 1975.

I. Csiszar, "Information-type measures of difference of probability distributions and indirect observations," *Studia Scientiarium Mathematicarum Hungarica*, vol. 2, pp. 299-318, 1967.

K. B. Davis, "Mean square error properties of density estimates," *Annals of Statistics*, vol. 5, pp. 1025-1030, 1975.

K. B. Davis, "Mean integrated square error properties of density estimates," *Annals of Statistics*, vol. 5, pp. 530-535, 1977.

P. Deheuvels, "Estimation non parametrique de la densite par histogrammes generalises," *Revue de Statistique Appliquee*, vol. 25, pp. 5-42, 1977.

L. Devroye, "On arbitrarily slow rates of global convergence in density estimation," *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 62, pp. 475-483, 1983.

L. Devroye, "The equivalence of weak, strong and complete convergence in L1 for kernel density estimates," *Annals of Statistics*, vol. 11, pp. 896-904, 1983.

L. Devroye, F. Machell, and C. S. Penrod, "The transformed kernel estimate," Technical Report, Applied Research Laboratories, The University of Texas at Austin, Austin, TX., 1983.

L. Devroye and C. S. Penrod, "Distribution-free lower bounds in density estimation," *Annals of Statistics*, vol. 12, pp. 1250-1262, 1984.

L. Devroye, "A note on the L1 consistency of variable kernel estimates," *Annals of Statistics*, vol. 13, pp. 1041-1049, 1985.

L. Devroye and L. Gyorfi, *Nonparametric Density Estimation. The L1 View*, John Wiley, New York, N.Y., 1985.

L. Devroye, "The kernel estimate is relatively stable," Technical Report, School of Computer Science, McGill University, 1986.

L. Devroye, "A universal lower bound for the kernel estimate," Technical Report, School of Computer Science, McGill University, 1986.

L. Devroye, "Nonparametric density estimates with improved performance on given sets of densities," Technical Report, School of Computer Science, McGill University, 1986.

R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis,* John Wiley, New York, N.Y., 1973.

B. Efron and C. Stein, "The jackknife estimate of variance," *Annals of Statistics,* vol. 9, pp. 586-596, 1981.

E. Fix and J. L. Hodges, "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties," Report 4, Project number 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX., 1951.

N. Glick, "Consistency conditions for probability estimators and integrals of density estimators," *Utilitas Mathematica,* vol. 6, pp. 61-74, 1974.

U. Grenander, "On the theory of mortality measurement, part II," *Skandinavisk Aktuarietidskrift,* vol. 39, pp. 125-153, 1956.

P. Groeneboom, "Estimating a monotone density," in *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II,* ed. L. Le Cam and R. A. Olshen, pp. 0-0, Wadsworth, Belmont, CA., 1983.

H. Guttmann and W. Wertz, "Note on estimating normal densities," *Sankhya, Series B,* vol. 38, pp. 231-236, 1976.

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

P. Hall, "Large-sample optimality of least squares cross-validation in density estimation," *Annals of Statistics,* vol. 11, pp. 1156-1174, 1983.

P. Hall, "Central limit theorem for integrated square error of multivariate nonparametric density estimators," *Journal of Multivariate Analysis,* vol. 14, pp. 1-16, 1984.

G. H. Hardy and W. W. Rogosinski, *Fourier Series,* Cambridge University Press, 1962.

J. A. Hartigan, *Clustering Algorithms,* John Wiley, New York, 1975.

J. A. Hartigan, "Consistency of single linkage for high-density clusters," *Journal of the American Statistical Association,* vol. 76, pp. 388-394, 1981.

W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association,* vol. 58, pp. 13-30, 1963.

T. Kawata, *Fourier Analysis in Probability Theory,* Academic Press, New York, 1972.

J. H. B. Kemperman, "On the optimum rate of transmitting information," in *Probability and Information Theory,* vol. 89, pp. 126-169, Springer-Verlag, Berlin, 1969.

A. N. Kolmogorov and V. M. Tikhomirov, "$\epsilon$-entropy and $\epsilon$-capacity of sets in function spaces," *Translations of the American Mathematical Society*, vol. 17, pp. 277-364, 1961.

S. Kullback, "A lower bound for discrimination information in terms of variation," *IEEE Transactions on Information Theory*, vol. 13, pp. 126-127, 1967.

L. LeCam, "Likelihood functions for large numbers of independent observations," in *Festschrift for Jerzy Neyman*, ed. F. N. Deavid, pp. 167-187, John Wiley, New York, 1966.

L. LeCam, "Convergence of estimates under dimensionality restrictions," *Annals of Statistics*, vol. 1, pp. 38-53, 1973.

D. O. Loftsgaarden and C. P. Quesenberry, "A nonparametric estimate of a multivariate density function," *Annals of Mathematical Statistics*, vol. 36, pp. 1049-1051, 1965.

G. G. Lorentz, *Approximation of Functions*, Holt, Rinehart and Winston, New York, 1965.

Ya. P. Lumelskii and P. N. Sapozhnikov, "Unbiased estimates of density functions," *Theory of Probability and its Applications*, vol. 14, pp. 357-364, 1969.

J. Marcinkiewicz and A. Zygmund, "Sur les fonctions independantes," *Fundamentales de Mathematiques*, vol. 29, pp. 60-90, 1937.

J. S. Marron, "An asymptotically efficient solution to the bandwidth problem of kernel density estimation," *Annals of Statistics*, vol. 13, pp. 1011-1023, 1985.

P. W. Millar, "Robust estimation via minimum distance methods," *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 55, pp. 73-89, 1981.

P. W. Millar, "A general approach to the optimality of minimum distance estimators," Technical Report, 1983.

E. Parzen, "On the estimation of a probability density function and the mode," *Annals of Mathematical Statistics*, vol. 33, pp. 1065-1076, 1962.

J. Pfanzagl, "On the existence of consistent estimates and tests," *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 10, pp. 43-62, 1968.

E. J. G. Pitman, *Some Basic Theory for Statistical Inference*, Chapman and Hall, London, 1979.

D. Pollard, "The minimum distance method of testing," *Metrika*, vol. 27, pp. 43-70, 1980.

M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *Annals of Mathematical Statistics*, vol. 27, pp. 832-837, 1956.

H. Scheffe, "A useful convergence theorem for probability distributions," *Annals of Mathematical Statistics*, vol. 18, pp. 434-458, 1947.

D. W. Scott and G. R. Terrell, "Biased and unbiased cross-validation in density estimation," Technical Report 23, Department of Statistics, Stanford University, 1986.

A. H. Seheult and C. P. Quesenberry, "On unbiased estimation of density functions," *Annals of Mathematical Statistics*, vol. 42, pp. 1434-1438, 1971.

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

H. Steinhaus and S. Kaczmarz, *Theorie der Orthogonalreihen,* Monografje Matematyczne VI, Warsaw, 1935.

C. J. Stone, "An asymptotically optimal window selection rule for kernel density estimates," *Annals of Statistics*, vol. 12, pp. 1285-1297, 1984.

C. J. Stone, "An asymptotically optimal histogram selection rule," in *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II,* ed. L. Le Cam and R. A. Olshen, pp. 513-520, Wadsworth, Belmont, CA., 1985.

G. S. Watson and M. R. Leadbetter, "On the estimation of the probability density," *Annals of Mathematical Statistics*, vol. 34, pp. 480-491, 1963.

E. Wegman, "Non-parametric probability density estimation I," *Technometrics*, vol. 14, pp. 533-546, 1972.

W. Wertz, "On unbiased density estimation," *An. Acad. Brasil. Cienc.*, vol. 47, pp. 65-72, 1975.

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

E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis,* Cambridge University Press, Cambridge, U.K., 1927.

P. Whittle, "Bounds for the moments of linear and quadratic forms in independent random variables," *Theory of Probability and its Applications*, vol. 5, pp. 302-305, 1960.

Y. G. Yatracos, "Rates of convergence of minimum distance estimators and Kolmogorov's entropy," *Annals of Statistics*, vol. 13, pp. 768-774, 1985.

# INDEX