Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

P. Groeneboom

Some current developments in density estimation

Department of Mathematical Statistics Report MS-R8503 June

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

Some current developments in density estimation

Piet Groeneboom
Centre for Mathematics and Computer Science
Amsterdam

Some recent results on the minimax risk of nonparametric density estimators, and, in particular, on the relation between minimax risk and metric entropy, are reviewed. We also discuss results on the distribution theory for the maximum likelihood estimator of a decreasing density and the connection of these results with properties of Brownian motion.

Key Words & Phrases: density estimation, minimax risk, metric entropy, smoothness restrictions, order restrictions, Grenander estimator, Brownian motion

AMS 1980 subject classifications: Primary 62G05, 62C20, 60F05, Secondary 60J65, 60J75.

1. Introduction

Every mathematician is probably familiar with the situation where he (she) is asked to describe to non-mathematicians the research he (she) is doing and to explain why this is an interesting and worthwhile endeavor. A very realistic description of what happens in such a case is given in the book by Davis and Hersh (1981), where on pp. 37-39 a "researcher on the decision problem for non-Riemannian hypersquares" is interviewed by a public information officer on the occasion of a renewal of his research grant.

A statistician who has to explain to a general mathematical public the kind of problems he (she) is interested in finds himself (herself) in a similar situation. In the following notes I will try to explain some current developments in the theory of (probability) density estimation in such a way that every mathematician should be able to understand it, and I apologize to statistical readers for the triviality of some of the remarks I will make. Furthermore, I have chosen for the approach of treating some typical examples in depth (with proofs), rather than covering a large area without really entering into the mathematics of the subject.

In statistical consultation, one is often confronted with the following problem. Someone (the client) shows graphs of a certain observed frequency distribution and asks "what theoretical probability distribution would fit this observed distribution?".

Figure 1.1 below shows an example of such a graph.

Report MS-R8503
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The netherlands

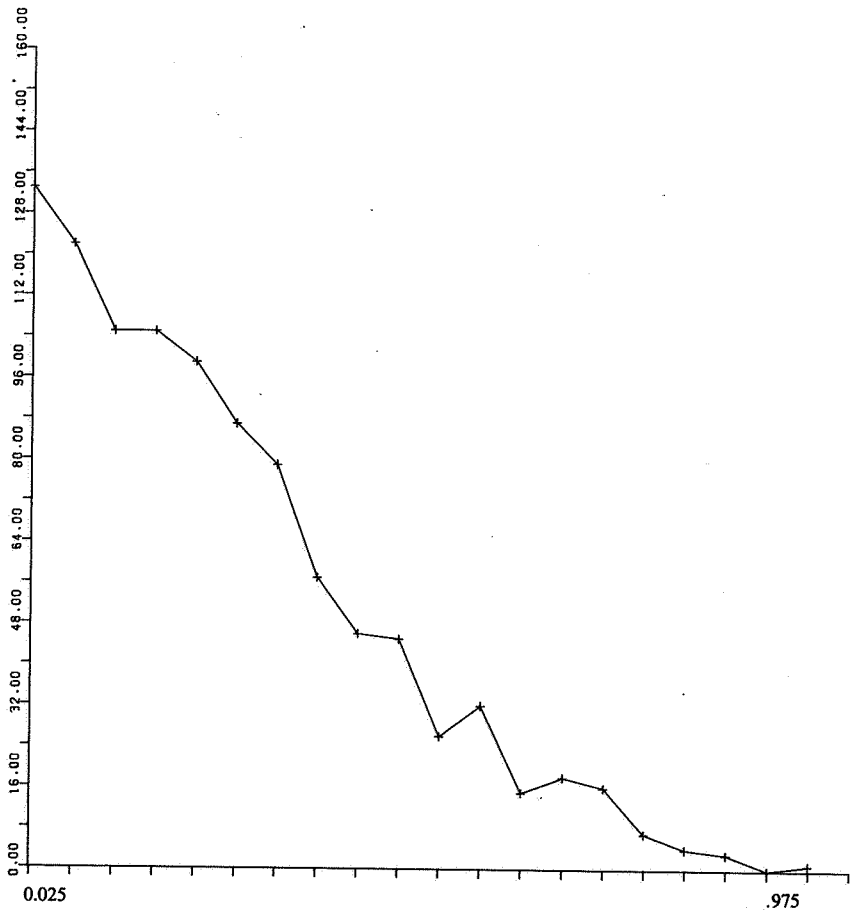


FIGURE 1.1. Frequency polygon, based on 100 observations, generated by a decreasing density on [0,1].

The graph is based on a sample of 1000 observations, generated by the STATAL random number generator from a decreasing density on [0,1] (to be specified later). The number of observations in each interval (of length 0.05), [0,.05), [.05,.1), etc. has been determined, and the graph connects linearly the values of these fractions (which are assigned to the midpoints of the intervals). This type of graph is called a *frequency polygon* and is familiar to everyone from the cartoons about worried businessmen looking at decreasing frequency polygons of sales figures.

Another method of summarizing these data is given by the *histogram* of figure 1.2. In this case one represents the fractions (or numbers) of observations in the intervals [0,.05) etc. by blocks where the height of the block indicates the fraction of observations in the interval.

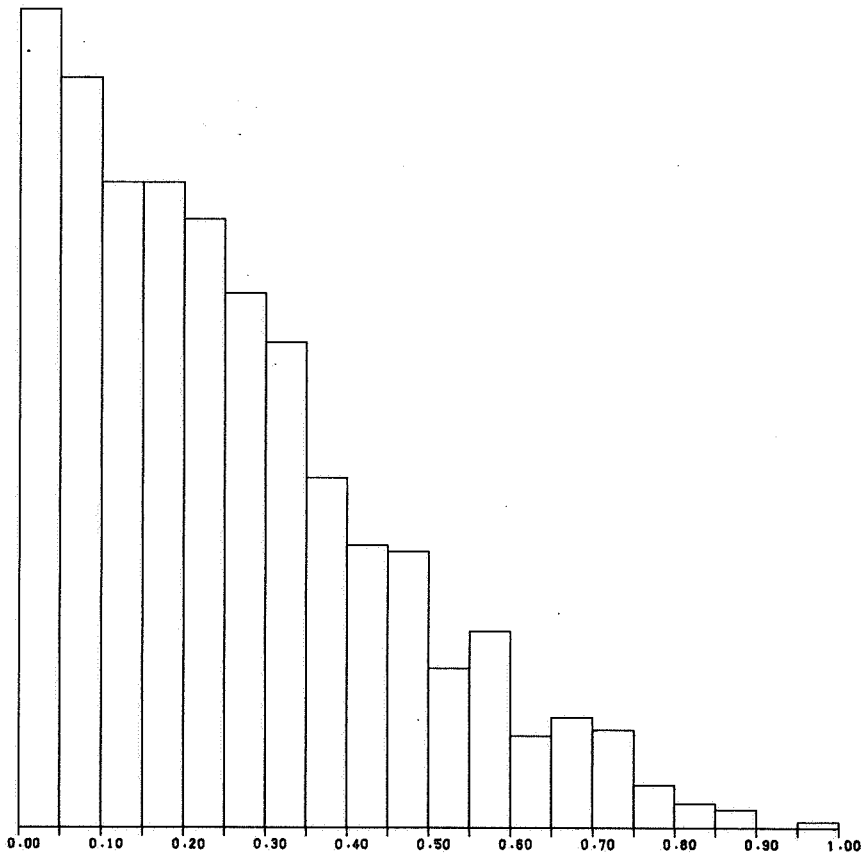
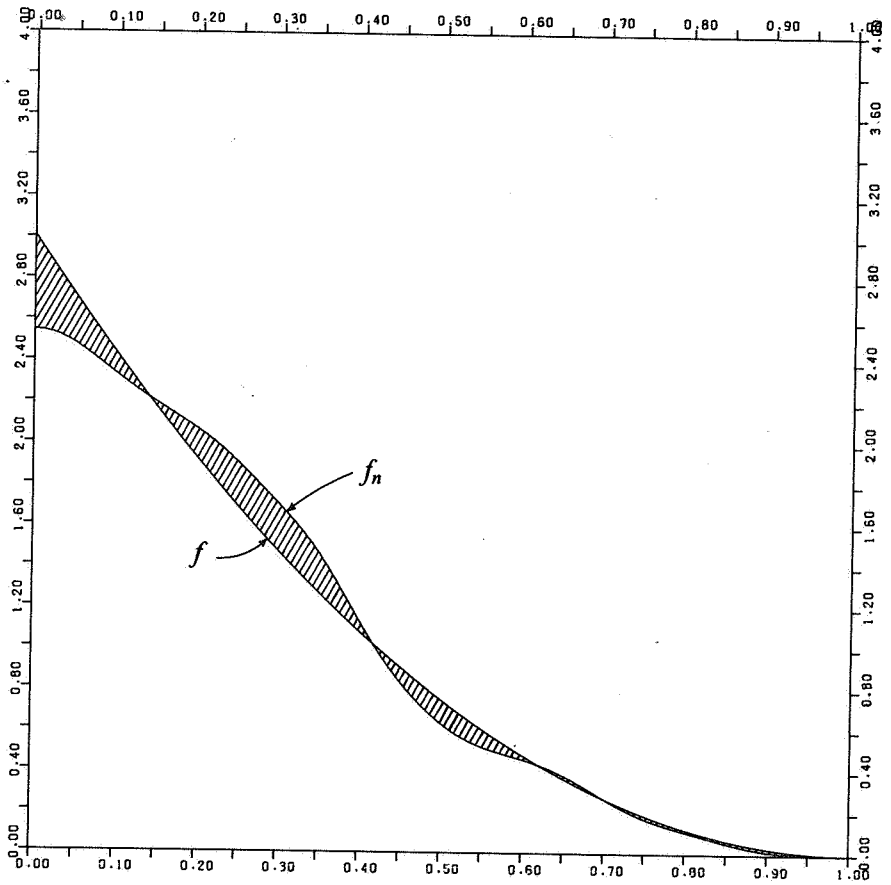


FIGURE 1.2. Histogram, based on 1000 observations, generated by a decreasing density on [0,1].

A third method of representing the observed distribution is given by *kernel estimators.* A kernel estimator of an unknown density f on [0,1], based on a random sample $X_1,...,X_n$ generated by the density f, is a function $f_{h,n}:[0,1]\to\mathbb{R}$ defined by

$$f_{h,n}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K((x-X_i)/h), \tag{1.1}$$

where the *kernel* K is a probability density (usually a Gaussian or "normal" density) and h is the *window size,* representing the degree of smoothing. Figure 1.3 shows a graph of a kernel estimate f_n, based on the same sample of $n = 1000$ observations that was used in figures 1.1 and 1.2. The density $f(x) = 3(1-x)^2$, $x\in[0,1]$, from which the observations were generated, is also shown in the graph. The area of the shaded region equals the L_1-distance between the kernel estimator and the density f.



$n=1000, f(t)=3(1-t)^2, t\in[0,1], K(x)=(2\pi)^{-\frac{1}{2}}\exp(-\frac{1}{2}x^2)$

FIGURE 1.3. Kernel estimate

Returning to the general question "What theoretical probability distribution would fit this empirical distribution (provided by the client)?", we can remark first of all that this question is meaningless if one does not specify beforehand

(i) the family $\mathcal{F}$ of densities one wishes to consider,

(ii) a *loss function* (usually a distance, such as the L_1 distance on $\mathcal{F}$), measuring the deviation between the real density and the estimator of the density.

In the same way, questions like "how big should the window size of my kernel estimator be?" or "how should I choose the intervals of my histogram?" are meaningless if (i) and (ii) above have not been specified.

In the old days, one only considered certain standard families of curves in the fitting problem, for example the Gaussian densities

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}, \quad x \in \mathbb{R} \tag{1.2}$$

which are completely specified by the two parameters μ and σ. But during the last two decades there has been an explosive development of techniques that are meant for the more general situation where one does not restrict the family of possible densities to a family parametrized by a subset of $\mathbb{R}^m$ ($m<\infty$), but instead considers infinite dimensional families. These techniques fall under the heading of *nonparametric density estimation* and have been greatly stimulated by developments in computer graphics.

In sections 2 and 3 we will give a discussion of the best performance one can expect from nonparametric density estimators according to the criterion of minimax risk. We will restrict ourselves to the choice of the L_1-distance as our loss function for densities on $\mathbb{R}^d (d<\infty)$. See for example figure 1.3, where $d=1$. This is a very natural loss function, since it corresponds to the total variation distance

$$D(P,Q) = \sup_{B\in\mathcal{B}} |P(B)-Q(B)| \tag{1.3}$$

between probability measures P and Q on $\mathbb{R}^d$, where $\mathcal{B}$ is the collection of Borel sets of $\mathbb{R}^d$. If P and Q are absolutely continuous with respect to Lebesque measure, with densities p and q respectively we have

$$\int_{\mathbb{R}^d} |p-q| = 2D(P,Q), \tag{1.4}$$

and unlike the L_2-distance for example, the L_1-distance is always well-defined and *invariant under monotone transformations of the coordinate axes* (a lucid account of the L_1-theory is given in the forthcoming book by L. Devroye and L. Györfi "Nonparametric density estimation: the L_1-view").

The minimax risk is defined as follows. Let $X_1,...,X_n$ be a random sample of n d-dimensional vectors, generated by a density f belonging to a class of densities $\mathcal{F}$ on $\mathbb{R}^d$. The *risk* under f of an estimator $\hat{f}_n = \hat{f}_n(.|X_1,...,X_n)$ of f, based on a sample $X_1,...,X_n$ from f, is the *expected value* of the L_1-distance between $\hat{f}_n$ and f:

$$E_f d_1(\hat{f}_n,f) = \int_{\mathbb{R}^{nd}}\cdots\int d_1(\hat{f}_n(.|x_1,...,x_n),f)\cdot f(x_1)...f(x_n)dx_1...dx_n \tag{1.5}$$

where d_1 denotes the L_1-distance and $x_i\in\mathbb{R}^d$, $1\leqslant i\leqslant n$. The *minimax risk* for the class $\mathcal{F}$, corresponding to samples of size n and loss function d_1, is now defined by

$$R_M(d_1,n) = \inf_{\hat{f}_n} \sup_{f\in\mathcal{F}} E_f d_1(\hat{f}_n,f) \tag{1.6}$$

where the infinum is taken over all possible density estimators $\hat{f}_n$ based on a random sample of n observations generated by a density in the family $\mathcal{F}$. Thus, a minimax estimator (if it exists), would minimize the maximum risk over all density estimators.

If one wants to estimate a parameter θ belonging to a finite-dimensional parameter set $\Theta\subset\mathbb{R}^m$ by an estimator $\hat{\theta}_n$ based on a sample of n observations, one usually has convergence of $\sqrt{n}(\hat{\theta}_n-\theta)$ to a limiting Gaussian distribution under the probability distribution specified by θ, as the sample size n tends to infinity. This means that the Euclidean distance between $\hat{\theta}_n$ and θ is of the order of $n^{-1/2}$ (the so-called "$\sqrt{n}$ law"). In nonparametric density estimation, the situation is radically different. The L_1-distance between a density f and its estimator $\hat{f}_n$, based on a sample of n observations generated by f, is typically of an order $n^{-\alpha}$, with $\alpha<1/2$.

In section 2 we will discuss the relation between the metric entropy (for the definition, see section 2) of the set of densities one wishes to consider and the rate of convergence to zero of the minimax risk. To our knowledge, this relation has for the first time been clearly pointed out by L. Birgé in his dissertation [4] (see also BIRGÉ (1983a)). Once this relation has been established, one can use results from approximation theory to give bounds and rates of convergence for the minimax risk. Roughly speaking, the bigger the metric entropy, the bigger the minimax risk (this relation can be exactly specified in certain cases, see Theorem 2.1). This is not surprising, since the metric entropy measures the "massivity" of a set, and the identifiability of a density within a set of densities will depend on the massivity of this set. Generally, (uniform) smoothness assumptions for the densities are reflected by the metric entropy of the set of densities: the smoother the densities are, the smaller the metric entropy will be (but we will give an example of a situation where things can go badly wrong, even for

a class of very smooth densities). Completely different types of restrictions can be put on the class of densities; for example, we may consider a class of decreasing densities on the interval [0,1], without any smoothness restrictions. The metric entropy of this set will again give us the rate of convergence to zero of the minimax risk. Hence, using the entropy concept, we can treat smoothness restrictions and order restrictions in a similar way.

In section 3 we give a fundamental lemma (Assouad's lemma), providing a lower bound for the minimax risk. We will also briefly discuss the concept of local asymptotic minimax risk, and give a local minimax result for the estimation of a monotone density (Theorem 3.1). Apart from this, the treatment of the minimax risk in sections 2 and 3 mainly uses the elegant techniques of BIRGÉ (1983a, 1983b, 1983c), with some simplifications which were made possible by the fact that we look at more special situations and do not try to obtain the best constants.

In section 4 we will discuss the behavior of a particular density estimator. We will also take a quick look at some distribution theory and the connection with Brownian motion.

2. METRIC ENTROPY AND UPPER BOUNDS FOR THE MINIMAX RISK

We first recall some definitions (see KOLMOGOROV and TIKHOMIROV (1961)). Suppose S is a subset of a metric space with metric d and let $\epsilon>0$. An $\epsilon-$*net* or $\epsilon-$*covering* of S is a subset $N \subset S$ such that

$$\forall s \in S \quad \exists n \in N: \quad d(n,s) \leqslant \epsilon \tag{2.1}$$

(often the ϵ in (2.1) is replaced by 2ϵ). A subset $A \subset S$ is called $\epsilon-$*separated* (on $\epsilon-$*distinguishable*) if

$$x,y \in A, \quad x \neq y \Rightarrow d(x,y) \geqslant \epsilon \tag{2.2}$$

Suppose S is totally bounded. Then, for each $\epsilon>0$, the (metric) $\epsilon-$*entropy* $H_\epsilon(S)$ of S is the logarithm of the *minimum* number of elements of an $\epsilon-$*net* of S. The $\epsilon-$*capacity* $C_\epsilon(S)$ of S is the logarithm of the *maximum* number of elements of an $\epsilon-$*separated* subset of S. The ϵ-entropy and ϵ-capacity satisfy the set of inequalities

$$C_{2\epsilon}(S) \leqslant H_{2\epsilon}(S) \leqslant C_\epsilon(S) \tag{2.3}$$

Suppose $\mathcal{F}$ is a set of probability densities on a compact set $S \subset \mathbb{R}^d$, matrized by the L_1-distance. Here and in the following, "density" will always mean "probability density with respect to Lebesgue measure". The following theorem specifies a relation between the behavior of the ϵ-entropy, as $\epsilon \downarrow 0$, and the rate of convergence to zero of the minimax risk as the sample size increases.

THEOREM 2.1. *Let $\mathcal{F}$ be a set of probability densities on a compact set $S \subset \mathbb{R}^d$, matrized by the L_1-distance d_1. Suppose that there exist numbers $\delta>0$ and $C_1>0$ such that for all $\epsilon>0$ the ϵ-entropy satisfies*

$$H_\epsilon(\mathcal{F}) \leqslant C_1 \epsilon^{-\delta} \tag{2.4}$$

Then there exist a constant $C_2>0$ such that

$$R_M(d_1,n) \leqslant C_2 n^{-1/(2+\delta)}, \quad n \in \mathbb{N}, \tag{2.5}$$

where $R_M(d_1,n)$ is the minimax risk for the class $\mathcal{F}$, corresponding to samples of size n and loss function d_1, defined by (1.6).

REMARK 2.1. The following result is (a part of) Theorem 1, Section 4 of Devroye and GYÖRFI (1985).

Let G be the set of densities on [0,1], bounded by $2+\delta$ (some $\delta>0$), and infinitely many times continuously differentiable on (0,1). Then we have for *any* sequence of density estimators $(\hat{f}_n)_n \in \mathbb{N}$, where $\hat{f}_n$ is based on a sample of size n,

(i)

$$\inf_n \sup_{f\in G} E_f \int |\hat{f}_n - f| \geqslant 1$$

(ii) For any sequence $(a_n)_n \in \mathbb{N}$ of positive numbers a_n tending to 0,

$$\sup_{f\in G} \limsup_{n\to\infty} a_n^{-1} E_f \int |\hat{f}_n - f| = \infty.$$

This result shows that conditions like compact support and smoothness are not sufficient to ensure a reasonable identifiability of the unknown density, but that we need a condition like (2.4) on the metric entropy of the class of densities, to have the risk of our estimators tend to zero uniformly for all densities in the class. No matter how sophisticated or "adaptive" our estimators $\hat{f}_n$ are, there will always be some densities in the class G which will be estimated rather poorly by this estimator.

Before giving the proof of Theorem 2.1, we give two examples of its use.

EXAMPLE 2.1. (Smooth densities) Suppose that $\mathcal{F}$ is the class of densities on [0,1] such that, for some $\alpha \in (0,1]$,

$$|f^{(p)}(x) - f^{(p)}(y)| \leqslant C|x-y|^\alpha, \quad x,y \in (0,1), \tag{2.6}$$

where $p \in \mathbb{N} \cup \{0\}$ and $C>0$ is a constant independent of f (i.e. condition (2.6) holds *uniformly* for $f \in \mathcal{F}$). Then there exists a constant $C_1>0$ such that the ϵ-entropy $H_\epsilon(\mathcal{F})$ satisfies

$$H_\epsilon(\mathcal{F}) \leqslant C_1 \epsilon^{-1/(p+\alpha)}, \tag{2.7}$$

and hence, by Theorem 2.1,

$$R_m(d_1,n) \leqslant C_2 n^{-(p+\alpha)/(1+2p+2\alpha)}, \tag{2.8}$$

for some constant $C_2>0$ and all $n \in \mathbb{N}$. Results like (2.7) can be found in papers on approximation theory, see e.g. KOLMOGOROV and TIKHOMIROV (1961) and LORENTZ (1966).

By the techniques that we will discuss in Section 3, it can be shown that there also exists a constant $C_3>0$ such that

$$R_M(d_1,n) \geqslant C_3 n^{-(p+\alpha)/(1+2p+2\alpha)} \tag{2.9}$$

for all $n \in \mathbb{N}$. Hence the "speed of estimation" in this density estimation problem is $n^{-(p+\alpha)/(1+2p+2\alpha)}$.

We now sketch the construction of an ϵ-net for the case $p=0$, i.e.

$$|f(x) - f(y)| \leqslant C|x-y|^\alpha, \quad x,y \in (0,1), \ \alpha \in (0,1], \tag{2.10}$$

for $f \in \mathcal{F}$ (f is *uniformly α-Hölder continuous*).

Fix $\epsilon>0$, let $\eta = 1/\{1+[(\epsilon/C)^{-1/\alpha}]\}$, where $[x]$ is the largest integer $\leqslant x$, and let $\mathcal{A}$ be the set of functions ϕ on [0,1] such that

$$\begin{cases} \phi(i\eta) = j\epsilon, \quad i,j \in \mathbb{N}\cup\{0\}, \quad i \leqslant \eta^{-1} \\ \phi((i+1)\eta) = \phi(i\eta) + k\epsilon, \quad k = -1,0, \quad \text{or} \quad 1 \\ \phi \quad \text{is linear on the intervals} \quad [i\eta,(i+1)\eta] \end{cases}$$

Figure 2.1 shows a picture of such a function ϕ on a part of the interval [0,1]

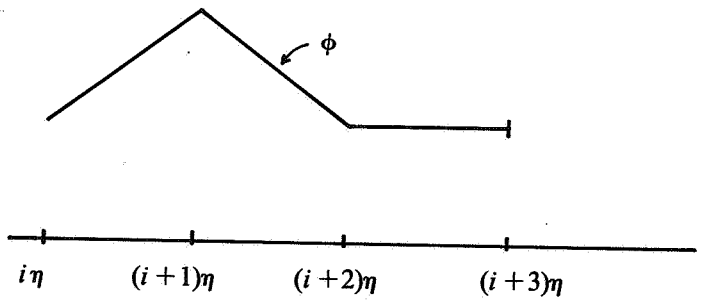


FIGURE 2.1

For each $f\in\mathcal{F}$ there exists $\phi\in\mathcal{A}$ such that $d_1(f,\phi)\leqslant 2\epsilon$. The set N of functions $\phi\in\mathcal{A}$ such that $d_1(f,\phi)\leqslant 2\epsilon$ for some $f\in\mathcal{F}$ is contained in the set of functions $\phi\in\mathcal{A}$ such that $1-\epsilon\leqslant\phi(i\eta)\leqslant 1+\epsilon$ for at least one index i (since $|f(i\eta)-1|\leqslant\epsilon$ for at least one index i, if $f\in\mathcal{F}$, using the fact that f is a probability density).

Hence $\mathrm{Card}(N)\leqslant(\eta+1)^{-1}3^{\eta^{-1}}$. Picking one density $f\in\mathcal{F}$ in each L_1-ball $B(\phi;2\epsilon)$ of radius 2ϵ around a ϕ such that $B(\phi;2\epsilon)\cap\mathcal{F}\neq\varnothing$ provides us with a 4ϵ-net $N_{4\epsilon}$ of $\mathcal{F}$ such that

$$\log\{\mathrm{Card}(N_{4\epsilon})\}\leqslant C_1\eta^{-1}=C'_1\epsilon^{-\frac{1}{\alpha}}$$

Thus there exists a constant $C'_1>0$ such that, for all $\epsilon>0$,

$$H_\epsilon(\mathcal{F})\leqslant C'_1\epsilon^{-1/\alpha}$$

For example, if $\alpha=1$ (a uniform Lipschitz condition on $\mathcal{F}$), we get $H_\epsilon(\mathcal{F})\leqslant C'_1\epsilon^{-1}$, and hence the speed of estimation is of order $n^{-1/3}$. We will meet the same speed of estimation in the next example on decreasing (but possibly discontinuous) densities on [0,1].

EXAMPLE 2.2. (Decreasing densities, BIRGÉ (1983c)). Suppose $\mathcal{F}$ is the family of decreasing (i.e. non-increasing) densities f on [0,1] such that $f\leqslant M$, for all $f\in\mathcal{F}$. We will show that there exists a $C_1>0$ such that the ϵ-entropy $H_\epsilon(\mathcal{F})$ satisfies

$$H_\epsilon(\mathcal{F})\leqslant C_1\epsilon^{-1},\quad \epsilon>0. \tag{2.11}$$

The following construction of a 4ϵ-net for $\mathcal{F}$ is based on BIRGÉ (1983c), with some simplifications due to the fact that we do not try to obtain the best (or at least a "very good") constant C_1.

Let $\epsilon\in(0,1)$ and $p\in\mathbb{N}$ satisfy

$$M=(1+\epsilon)^p-1. \tag{2.12}$$

To avoid trivialities, we suppose $M>1$ (otherwise $\mathcal{F}$ only consists of one element: the uniform density on [0,1]). Define, for $0\leqslant i\leqslant p$

$$x_i=M^{-1}\{(1+\epsilon)^i-1\},\quad y_i=(1+\epsilon)^i-1, \tag{2.13}$$

and for $0<i\leqslant p$,

$$I_i=[x_{i-1},x_i) \tag{2.14}$$

The length l_i of the i-th interval I_i is $M^{-1}\epsilon(1+\epsilon)^{i-1}$. The 4ϵ-net that we will construct, will be based on the finite set $\mathcal{G}$ of functions g, which are constant on the intervals I_i and take values in the set

$$Y=\{y_0,\ldots,y_p\}. \tag{2.15}$$

Now let f be a decreasing density on [0,1]. We define

$$f_i=f(x_i);\quad \bar{f}_i=l_i^{-1}\int_{I_i}f(x)dx. \tag{2.16}$$

Then $\Sigma_{i=1}^{p} l_i f_i \leq \Sigma_{i=1}^{p} l_i \bar{f}_i = 1$. Suppose

$$\bar{f}_i = \lambda y_{j-1} + (1-\lambda) y_j \qquad \begin{cases} 0 \leq \lambda \leq 1 \\ y_{j-1}, y_j \in Y \end{cases} \tag{2.17}$$

(see figure 2.2).

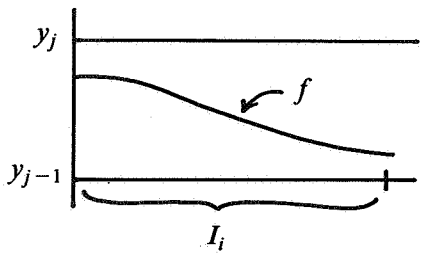


FIGURE 2.2.

Then we define the approximating function g on I_i by

$$g = \begin{cases} y_{j-1}, \text{ if } \lambda > \frac{1}{2} \\ \\ y_j \ , \text{ if } \lambda \leq \frac{1}{2} \end{cases} \tag{2.18}$$

The function g is decreasing, non-negative, and $g \leq M$. If $\lambda \leq \frac{1}{2}$, we get on the interval I_i

$$|\bar{f}_i - g| = |\bar{f}_i - y_j| \leq \tfrac{1}{2}\epsilon(1+\bar{f}_i) \tag{2.19}$$

and similarly $|\bar{f}_i - g| \leq \frac{1}{2}\epsilon(1+\bar{f}_i)$, if $\lambda > \frac{1}{2}$. Hence

$$\int_{I_i} |\bar{f}_i - g| dx \leq \tfrac{1}{2}\epsilon l_i (1+\bar{f}_i). \tag{2.20}$$

Since

$$\int_{I_i} |f(x) - \bar{f}_i| dx \leq \tfrac{1}{2} l_i (f_{i-1} - f_i)$$

and $l_{i+1} = (1+\epsilon) l_i$, we now get

$$\begin{aligned} \int_0^1 |f(x) - g(x)| dx &\leq \sum_{i=1}^{p} \{ \int_{I_i} |f(x) - \bar{f}_i| dx + \int_{I_i} |\bar{f}_i - g| dx \} \\ &\leq \tfrac{1}{2} \{ l_1 f_0 + \epsilon \sum_{i=1}^{p-1} l_i f_i \} + \epsilon \leq 2\epsilon \end{aligned}$$

Thus, for each decreasing density f on [0,1], such that $f \leq M$, there exists a decreasing function g, which is constant on the intervals I_i and takes values in the set $Y = \{y_0, \ldots, y_p\}$, satisfying

$$\int_0^1 |f(x) - g(x)| dx \leq 2\epsilon \tag{2.21}$$

The number of functions g of this type equals the number of ways one can choose p non-negative integers k_j such that $\Sigma_{j=1}^{p} k_j \leq p$. This number in turn equals the number of ways we can pick $p+1$

non-negative integers $k_1,\ldots,k_{p+1}$ such that $\Sigma_{j=1}^{p+1} k_j = p$. This number is $\binom{2p}{p}$ (see e.g. Feller (1968), Section II. 5)

Choosing one $f \in \mathcal{F}$ for each $g \in \mathcal{G}$ in such a way that $d_1(f,g) \leqslant 2\epsilon$, provides us with a 4ϵ-net $N_{4\epsilon}$ of $\mathcal{F}$ such that

$$\text{Card}(N_{4\epsilon}) \leqslant \binom{2p}{p}$$

Since $\binom{2p}{p} \leqslant \frac{2^{2p}}{\sqrt{\pi p}}$, we have

$$H_{4\epsilon}(\mathcal{F}) \leqslant 2p \log 2 = (2 \log 2) \cdot \frac{\log(M+1)}{\log(1+\epsilon)}$$

and $1/\log(1+\epsilon) \sim \epsilon^{-1}$, as $\epsilon \downarrow 0$. Thus there exists a constant $C_1 > 0$ such that: for all $\epsilon > 0$,

$$H_\epsilon(\mathcal{F}) \leqslant C_1 \epsilon^{-1}$$

and hence, by Theorem 2.1,

$$R_M(d_1,n) \leqslant C_2 n^{-1/3},$$

for some constant $C_2 > 0$. We will show in Section 3 that there also exists a constant $C_3 > 0$ such that

$$R_M(d_1,n) \geqslant C_3 n^{-1/3}, \quad n \in \mathbb{N},$$

implying that the speed of estimation is of order $n^{-1/3}$ for this estimation problem.

The proof of theorem 2.1 is based on the felicitous idea, introduced by Le Cam and further developed in the context of density estimation by BIRGÉ (1980, 1983), of constructing estimators on the basis of a family of tests of hypotheses. Birgé calls these estimators *"d-estimators"*, where d denotes the distance function, used to define the loss-function (in our case the L_1-distance). These estimators are concentrated on a ϵ-net and they give a connection between the ϵ-entropy and the minimax risk.

We now give the construction of the d-estimators based on one observation generated by a probability distribution p_θ, where θ belongs to a parameter set Θ. Suppose Θ is metrized by a metric d and totally bounded for this metric. Let, for $\epsilon > 0$, N_ϵ be an ϵ-net of Θ and $\{B_s,\ s \in N_\epsilon\}$ be the family of balls, with radius ϵ and centers $s \in N_\epsilon$, covering Θ. Furthermore, let $\{\phi_{s,t}\}$ be a family of tests $\phi_{s,t}$ between the balls B_s and B_t for parameters $s,t \in N_\epsilon$, $s \neq t$, and let $J_s(X)$ be the set of t's in N_ϵ such that the tests $\phi_{s,t}$ rejects B_s and accepts B_t on the basis of the observation X. We suppose that $\phi_{s,t}$ is a real-valued function, defined on the space of possible observations, such that $0 \leqslant \phi_{s,t} \leqslant 1$, and

$$\begin{cases} \phi_{s,t}(X) = 1 \Leftrightarrow B_s \text{ is rejected and } B_t \text{ is accepted} \\ \phi_{s,t}(X) = 0 \Leftrightarrow B_s \text{ is accepted and } B_t \text{ is rejected} \end{cases} \tag{2.22}$$

Define

$$L_s(X) = \begin{cases} \max\limits_{t \in J_s(X)} d(s,t) \\ \\ 0, \text{ if } J_s(X) = \varnothing \end{cases} \tag{2.23}$$

A d-estimator is now defined in the following way

DEFINITION 2.1. A *d-estimator*, based on the family of tests $\{\phi_{s,t}\}$ is a point $\hat{\theta}(X) = t \in N_\epsilon$ such that

$$L_t(X) = \min_{s \in N_\epsilon} L_s(X). \tag{2.24}$$

In other words: a d-estimator is a point $s \in N_\epsilon$ such that the maximal distance $d(s,t)$ to "preferred" points $t \in N_\epsilon$ (i.e. points t such that $\phi_{s,t}(X) = 1$) is minimized.

Now let $X_1,...,X_n$ be a sample, generated by a density $f \in \mathcal{F}$, where $\mathcal{F}$ is as in Theorem 2.1, metrized by the L_1-distance d_1. A sample $X = (X_1,...,X_n)$ can be considered as *one* observation, generated by the product measure P_f^n, where P_f is the probability measure corresponding to f. We identify $\mathcal{F}$ with the set $\Theta = \{P_f^n : f \in \mathcal{F}\}$ and metrize Θ by $d(P_f^n, P_g^n) = d_1(f,g)$.

Let $B(f;\epsilon) = \{h \in \mathcal{F}: d_1(h,f) \leqslant \epsilon\}$ and $B(g;\epsilon) = \{h \in \mathcal{F}: d_1(h,g) \leqslant \epsilon\}$ be two closed ϵ balls in $\mathcal{F}$. In the problem of testing a ball $B(f;\epsilon)$ against a ball $B(g;\epsilon)$ we call a type I error, the error of concluding that the observations were generated by a density $h \in B(g;\epsilon)$, whereas they were actually generated by a density $h \in B(f;\epsilon)$, and a type II error, the error of concluding that the observations were generated by a density $h \in B(f;\epsilon)$ whereas they were actually generated by a density $h \in B(g;\epsilon)$. The following lemma shows that the probabilities of type I and type II errors tend to zero exponentially fast as the sample size increases (if the balls $B(f;\epsilon)$ and $B(g;\epsilon)$ are disjoint).

LEMMA 2.1. *There exists a test $\phi_{f,g}$ between $B(f;\epsilon)$ and $B(g;\epsilon)$, based on a sample of size n, such that the sum of the maximal probabilities of errors of the first and the second kind is dominated by α_n, where*

$$\alpha_n = \exp\{-\tfrac{1}{8}n(d_1(f,g)-2\epsilon)_+^2\} \tag{2.25}$$

Here $x_+ = x$, if $x \geqslant 0$, and 0 otherwise. Otherwise stated:

$$\sup_{h \in B(f;\epsilon)} \int \phi_{f,g} dP_h^n + \sup_{h \in B(g;\epsilon)} \int (1-\phi_{f,g}) dP_h^n \leqslant \alpha_n \tag{2.26}$$

SKETCH OF PROOF. Let $D(P,Q) = \sup_{B \in \mathcal{B}} |P(B)-Q(B)|$ be the total variation distance between two probability measures on $\mathbb{R}^d$, where $\mathcal{B}$ is the collection of Borel sets of $\mathbb{R}^d$ (generated by the Euclidean topology). If P and Q have densities p and q w.r.t. Lebesgue measure, we have

$$D(P,Q) = \tfrac{1}{2}d_1(p,q),$$

(see (1.3) and (1.4)).

Let $\mathcal{P}$ be the set of probability measures on $\mathbb{R}^d$ (or, more correctly, on $\mathcal{B}$) and let

$$\begin{cases} B_1 = \{P \in \mathcal{P} : 2D(P,P_f) \leqslant \epsilon\} \\ B_2 = \{P \in \mathcal{P} : 2D(P,P_g) \leqslant \epsilon\}. \end{cases} \tag{2.27}$$

Then $B(f;\epsilon) \subset B_1$ and $B(g;\epsilon) \subset B_2$ and the distance between the balls B_1 and B_2 is $\frac{1}{2}(d_1(f,g)-2\epsilon)_+$, if we use the total variation distance D. The balls B_1 and B_2 are convex and weakly compact *(unlike the L_1-balls $B(f;\epsilon)$ and $B(g;\epsilon)$)*. It now follows that

$$\nu_0 = \sup\{P: P \in B_1\}$$

is a 2-*alternating capacity* (CHOQUET 1953-1954, 1959) and that

$$\nu_1 = \inf\{P: P \in B_2\}$$

is a 2-*monotone capacity,* see HUBER and STRASSEN (1973) and BEDNARSKI (1982). Let $d_1(f,g)-2\epsilon > 0$. It then follows from the results in the last-mentioned papers that there exists a mutually absolutely continuous pair (P,Q), with $P \in B_1$ and $Q \in B_2$, and a test ϕ of the form

$$\phi(x_1,...,x_n) = \begin{cases} 1, & \text{if } \prod_{i=1}^n \frac{dQ}{dP}(x_i) \geqslant 1 \\ 0, & \text{otherwise,} \end{cases} \tag{2.28}$$

such that

$$\int \phi dP^n + \int (1-\phi) dQ^n = \sup_{P_1 \in B_1} \int \phi dP_1^n + \sup_{P_2 \in B_2} \int (1-\phi) dP_2^n$$

$$= 1 - \inf\{D(P_1^n, P_2^n) : P_1 \in B_1,\ P_2 \in B_2\}.$$

Such a pair (P,Q) is called a *least favorable pair* for testing B_1 against B_2 (or a *least informative experiment* in the terminology of LE CAM (1972) and BEDNARSKI (1982)).

Take $\phi_{f,g} = \phi$ and put $\mu = P + Q$. Then it can be shown that

$$\int \phi_{f,g} dP^n + \int (1-\phi_{f,g}) dQ^n = 1 - D(P^n, Q^n)$$

$$\leqslant exp\{-\frac{n}{2} \int \{(\frac{dP}{d\mu})^{1/2} - (\frac{dQ}{d\mu})^{1/2}\}^2 d\mu\}$$

$$\leqslant exp\{-\tfrac{1}{8} n (d_1(f,g) - 2\epsilon)^2\}. \quad \square$$

REMARK 2.2. We note that the least favorable pair of probability measures (P,Q), introduced in the proof of Lemma 2.1, does not necessarily consist of probability measures which are absolutely continuous w.r.t. Lebesgue measure. Thus the test of $B(f;\epsilon)$ against $B(g;\epsilon)$, satisfying (2.26), may be based on a pair of probability measure *outside these balls.* This is caused by the fact that generally $B(f;\epsilon)$ and $B(g;\epsilon)$ are not weakly compact.

Proof of Theorem 2.1.

Fix $\epsilon > 0$, and choose an ϵ-net N_ϵ of $\mathcal{F}$ such that

$$\log\{\text{Card}(N_\epsilon)\} \leqslant C_1 \epsilon^{-\delta} \tag{2.29}$$

(see condition (2.4) of Theorem 2.1). By Lemma 2.1 there exists a family of tests $\{\phi_{f,g}\}$, based on samples of size n, where $f, g \in N_\epsilon$ and $\phi_{f,g}$ is a test between the balls $B(f;\epsilon)$ and $B(g;\epsilon)$ satisfying (2.26).

Let $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ be a d-estimator, based on the family of tests $\{\phi_{f,g}\}$ (see Definition 2.1.). Fix $f \in \mathcal{F}$ and let $g \in N_\epsilon$ be a density such that $d_1(f,g) \leqslant \epsilon$. Furthermore, let N_i be the number of densities $h \in N_\epsilon$ such that $(i+2)\epsilon \leqslant d_1(h,g) < (i+3)\epsilon$. Then we have, by Lemma 2.1, for $i \geqslant 1$,

$$P_f\{d_1(\hat{\theta}_n, f) \geqslant (3+i)\epsilon\} \leqslant P_f\{d_1(\hat{\theta}_n, g) \geqslant (2+i)\epsilon\}$$

$$\leqslant \sum_{j \geqslant i} N_j exp\{-\tfrac{1}{8} n j^2 \epsilon^2\}$$

Hence,

$$E_f d_1(\hat{\theta}_n, f) \leqslant 4\epsilon + \epsilon \sum_{i \geqslant 1} P_f\{d_1(\hat{\theta}_n, f) \geqslant (3+i)\epsilon\}$$

$$\leqslant \epsilon(4 + \sum_{i \geqslant 1} \sum_{j \geqslant i} N_j exp\{-\tfrac{1}{8} n j^2 \epsilon^2\}$$

$$= \epsilon(4 + \sum_{i \geqslant 1} i N_i exp\{-\tfrac{1}{8} n i^2 \epsilon^2\})$$

Let $n \geqslant 8C_1$, where C_1 is as in (2.29), and choose $\epsilon > 0$ in such a way that $\epsilon^{2+\delta} = 8C_1/n$. Then the function $j \to j\ exp\{-\frac{1}{8} n j^2 \epsilon^2\}$ is decreasing for $j \geqslant j_0 = 1 + [1/C_1]$, and hence

$$E_f d_1(\hat{\theta}_n, f) \leqslant \epsilon(4 + \sum_{i \geqslant 1} i N_i exp\{-\tfrac{1}{8} n i^2 \epsilon^2\})$$

$$\leqslant 4\epsilon + \epsilon j_0 \text{Card}(N_\epsilon) exp\{-\tfrac{1}{8} n \epsilon^2\}$$

$$= (4+j_0)(8C_1)^{1/(2+\delta)} n^{-1/(2+\delta)}$$

Put $C=(4+j_0)(8C_1)^{1/(2+\delta)}$. Then we get $\sup_{f\in\mathfrak{F}} E_f d_1(\hat{\theta}_n, f) \leqslant Cn^{-1/(2+\delta)}$. □

3. Lower bounds for the minimax risk

In obtaining lower bounds for the minimax risk, we compare two kinds of "distinguishability" of the probability densities:

1) distinguishability in terms of the loss function on the set of densities
2) distinguishability in terms of some "information measure".

Usually the distinguishability in terms of an information measure is measured by the *Kullback-Leibler information*

$$K(Q,P) = \begin{cases} \int \log\left[\frac{dQ}{dP}\right] dQ, & \text{if } Q \ll P \\ \infty, & \text{otherwise} \end{cases} \tag{3.1}$$

One then uses an information-theoretic lemma *(Fano's Lemma,* see e.g. Ibragimov and Hasminskii (1981), p. 323-325) to give lower bounds for the minimax risk. This technique is used in e.g. Birgé (1980, 1983), Bretagnolle and Huber (1979) and Ibragimov and Hasminskii (1980, 1981).

Here we give another Lemma (Assouad's Lemma), where the Kullback-Leibler information $K(Q,P)$ is replaced by the *Hellinger distance* $h(P,Q)$, defined by

$$h(P,Q) = \{\tfrac{1}{2}\int \{\left[\frac{dP}{d\mu}\right]^{1/2} - \left[\frac{dQ}{d\mu}\right]^{1/2}\}^2 d\mu\}^{1/2}, \tag{3.2}$$

where μ is any measure dominating P and Q (for example: $\mu = P+Q$). Roughly speaking, the Hellinger distance can be considered as a local version of the Kullback-Leibler information; it has the advantage of being a *distance.* The Kullback-Leibler information, sometimes called "Kullback-Leibler distance" is not really a distance (it does not satisfy the triangle inequality).

We now state Assouad's Lemma in a form slightly adapted to our purposes.

Lemma 3.1 (Assouad, 1982). *Let $A_r=\{0,1\}^r = \{a:a=(a_1,...,a_r),\ a_i=0 \text{ or } 1\}$ and let $\mathfrak{F}$ be a collection of probability densities on $\mathbb{R}^d$. Suppose that $\phi: a \to f_a$ is a bijection of A_r on a subset $\mathfrak{F}_r$ of $\mathfrak{F}$ and that $\{B_1,...,B_r\}$, is a partition of $\mathbb{R}^d$ into measurable sets $B_1,...,B_r$ such that, if $a=(a_1,...,a_{i-1}, 1, a_{i+1},...,a_r)$ and $a'=(a_1,...,a_{i-1}, 0, a_{i+1},...,a_r)$*

$$\frac{1}{2}\int_{\mathbb{R}^d} (f_a^{1/2} - f_{a'}^{1/2})^2 dx \leqslant \beta_i \leqslant 1 \tag{3.3}$$

$$\int_{B_i} |f_a - f_{a'}| dx \geqslant \alpha_i > 0. \tag{3.4}$$

Let $\hat{f}_n$ be any *density estimator, based on a sample of size n, generated by a density $f\in\mathfrak{F}$. Then*

$$\sup_{f\in\mathfrak{F}} E_f \int_{\mathbb{R}^d} |\hat{f}_n - f| dx \geqslant \frac{1}{2}\sum_{i=1}^{r} \alpha_i \max\{1-(2n\beta_i)^{1/2},\ \tfrac{1}{2}(1-\beta_i)^{2n}\}.$$

We omit the proof of this lemma, but instead discuss some applications. Usually the α_i's are taken equal to some α and the β_i's taken equal to some β. One then looks for densities which are α-separated in the L_1-distance, but have the smallest possible Hellinger distance. In fact we are then dealing with the (local) α-capacity of the set $\mathfrak{F}$. Hopefully these remarks will become more clear by

looking at some examples.

EXAMPLE 3.1 (Continuation of Example 2.1). Suppose that $\mathcal{F}$ consists of the densities f on [0,1], satisfying

$$|f(x)-f(y)| \leqslant C|x-y|^{\alpha}, \quad x,y \in [0,1]$$

where C is independent of f. Let $\epsilon \in (0,1)$ and $\eta = \{1+[(\frac{1}{4}\epsilon/C)^{-1/\alpha}]^{-1}$. Suppose $b_j = j\eta \leqslant 1$, for some positive integer j and let f_j be the piecewise linear function defined on $[b_{j-1}, b_j]$ by $f_j(b_{j-1})=0$, $f_j(b_{j-1}+\frac{1}{4}\eta)=\frac{1}{4}\epsilon$, $f_j(b_{j-1}+\frac{1}{2}\eta)=0$, $f_j(b_{j-1}+\frac{3}{4}\eta)=-\frac{1}{4}\epsilon$, $f_j(b_j)=1$, and f_j is linearly interpolated for other values $x \in [b_{j-1}, b_j]$ (see figure 3.1)

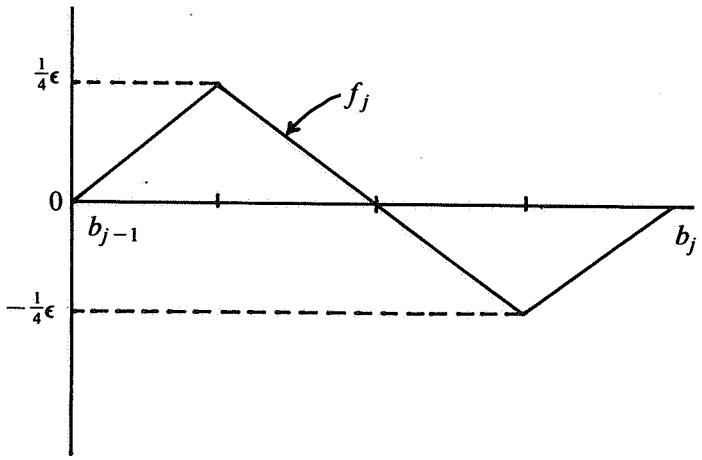


FIGURE 3.1.

Define $r=\eta^{-1}$ and $\mathcal{F}_r=\{f:[0,1]\to\mathbb{R} \mid f=1+\Sigma_{i=1}^{r}\lambda_i f_i,\ \lambda_i=\pm 1\}$. Let the function $\phi:\{0,1\}^r \to \mathcal{F}_r$ be defined by

$$\phi(a) = 1+\sum_{i=1}^{r}\lambda_i f_i, \quad \begin{cases} \lambda_i = 1, & \text{if } a_i = 1 \\ \lambda_i = -1, & \text{if } a_i = 0. \end{cases}$$

Then, if $a=(a_1,\ldots,a_{i-1},\ 1,\ a_{i+1},\ldots,a_r)$ and $a'=(a_1,\ldots,a_{i-1},\ 0,\ a_{i+1},\ldots,a_r)$

$$\int_{b_{i-1}}^{b_i} |f_a - f_{a'}|dx = \frac{1}{4}\epsilon\eta, \tag{3.5}$$

and the Hellinger distance satisfies

$$h^2(f_a, f_{a'}) \leqslant \frac{1}{12}\eta\epsilon^2 \tag{3.6}$$

Thus the conditions (3.3) and (3.4) of Assouad's Lemma are satisfied, with $\alpha_i = 1/4\epsilon\eta$ and $\beta_i = (1/12)\eta\epsilon^2$, and we obtain for $n=[1/(\eta\epsilon^2)]$ and *any* density estimator $\hat{f}_n$ based on a sample of size n generated by a density $f \in \mathcal{F}$,

$$\sup_{f\in\mathcal{F}} E_f \int_{\mathbb{R}} |\hat{f}_n - f|dx \geqslant \epsilon(1-6^{-\frac{1}{2}})/8 \tag{3.7}$$

Hence the minimax risk $R_M(d_1,n)$ satisfies (for a constant $c>0$)

$$R_M(d_1,n) \geqslant c \,.\, n^{-\alpha/(1+2\alpha)}$$

Since, by Example 2.1 ((2.8) with $p=0$),

$$R_M(d_1,n) \leqslant C' \,.\, n^{-\alpha/(1+2\alpha)}$$

for some $C'(>c)$, the speed of convergence to zero of the minimax risk is of order $n^{-\alpha/(1+2\alpha)}$.

EXAMPLE 3.2 (Continuation of example 2.2). Let $\mathcal{F}$ be the set of decreasing densities on [0,1], such that $f \leqslant M$ for each $f \in \mathcal{F}$, with $M>1$. We will show that the minimax risk satisfies

$$R_M(d_1,n) \geqslant Cn^{-\frac{1}{3}}, \quad n \in \mathbb{N} \tag{3.8}$$

Since it was shown in section 2 that $R_M(d_1,n) \leqslant C' n^{-1/3}$, for all $n \in \mathbb{N}$ and some $C'>0$, the minimax risk tends to zero at the rate $n^{-1/3}$.

Let $\epsilon \in (0, \frac{1}{2})$, $r \in \mathbb{N}$, $u = \{(1+\epsilon)^r - 1\}^{-1}$, $\lambda = (1+\epsilon)/\{ru\epsilon(1+\frac{1}{2}\epsilon)\}$, and $x_i = u\{(1+\epsilon)^i - 1\}$, $0 \leqslant i \leqslant r$.

Define, for $1 \leqslant i \leqslant r$, the intervals I_i by $I_i = [x_{i-1}, x_i)$. The interval I_i has length $l_i = u\epsilon(1+\epsilon)^{i-1}$. Let the functions f_i and g_i be defined on the interval I_i by

$$f_i(x) = \lambda(1+\epsilon)^{-i}(1+\tfrac{1}{2}\epsilon), \quad x \in I_i, \tag{3.9}$$

and

$$g_i(x) = \begin{cases} \lambda(1+\epsilon)^{-i+1}, & \text{first half of } I_i \\ \lambda(1+\epsilon)^{-i}, & \text{second half of } I_i \end{cases} \tag{3.10}$$

(see figure 3.2).

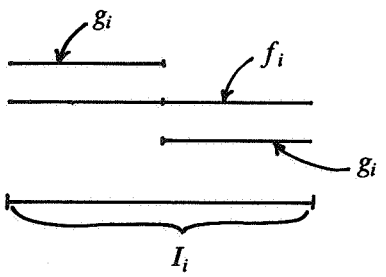


FIGURE 3.2

Then $\int_{I_i} g_i dx = \int_{I_i} f_i dx = 1/r$, and

$$\int_{I_i} |f_i - g_i| dx = \epsilon/\{(2+\epsilon)r\}, \tag{3.11}$$

$$\frac{1}{2} \int_{I_i} \{\sqrt{f_i} - \sqrt{g_i}\}^2 dx < \frac{1}{32}\epsilon^2/r. \tag{3.12}$$

Now let $\mathcal{F}_r$ be the family of 2^r functions, defined on [0,1) by

$$f = \sum_{i=1}^{r} (\lambda_i f_i + (1-\lambda_i) g_i) 1_{I_i}, \tag{3.13}$$

where $\lambda_i = 0$ or 1 and 1_{I_i} is the indicator function of the interval I_i. Suppose that ϵ and r satisfy

$$g_1(0) = \frac{1+\epsilon}{r\epsilon(1+\frac{1}{2}\epsilon)} \{(1+\epsilon)^r - 1\} \leqslant M. \tag{3.14}$$

Since $f \leqslant g_1(0)$, if $f \in \mathcal{F}_r$, we have $\mathcal{F}_r \subset \mathcal{F}$ if (3.14) is satisfied. Hence, by (3.11), (3.12) and Lemma 3.1,

$$R_M(d_1,n) \geqslant \tfrac{1}{2}\frac{\epsilon}{2+\epsilon}\{1-\sqrt{\frac{2n\epsilon^2}{32r}}\} \tag{3.15}$$

Choose, for each $r \in \mathbb{N}$, the number $\epsilon_r > 0$ such that $(1+\epsilon_r)^r = M$. Then

$$(1+\epsilon_r)\{(1+\epsilon_r)^r - 1\}/\{r\epsilon_r(1+\tfrac{1}{2}\epsilon_r)\} \sim \frac{M-1}{\log M}, \tag{3.16}$$

as $r \to \infty$. Since $(M-1)/\log M < M$, for $M > 1$, there exists r_0 such that the left-hand side of (3.16) is smaller than M, if $r \geqslant r_0$. Hence, by (3.15), $\mathcal{F}_r \subset \mathcal{F}$, if $r \geqslant r_0$. Taking $n = [r/\epsilon_r^2]$ yields

$$R_M(d_1,n) \geqslant \frac{\epsilon_r}{2(2+\epsilon_r)}(1-\tfrac{1}{4}) \sim (3/16)\,(\log M)^{1/3} n^{-1/3},$$

as $r \to \infty$ (and hence $\epsilon_r \downarrow 0$, $n \to \infty$). Thus there exists a constant $C > 0$ such that (3.8) holds.

REMARK 3.1. BIRGÉ (1983c) gives a better constant in the lower bound of the minimax risk (at the cost of more difficult computations).

REMARK 3.2. The restriction to the interval [0,1] in Examples 3.1 and 3.2 is not essential, but the restriction to compact intervals is. For example, if $\mathcal{F}$ is the family of decreasing densities on $[0,\infty)$, we get arbitrarily slow rates of convergence for the minimax risk (like in Remark 2.1), even if $f \leqslant M$, for all $f \in \mathcal{F}$.

If $\mathcal{F}$ is the family of decreasing densities f on $[0,L]$ such that $f \leqslant M$, for all $f \in \mathcal{F}$, we obtain by similar computations as in examples 2.2 and 3.2

$$C_1(\log LM)^{1/3} n^{-1/3} \leqslant R_M(d_1,n) \leqslant C_2(\log LM)^{1/3} n^{-1/3} \tag{3.17}$$

Hence, for fixed n, the minimax risk grows at the rate $(\log LM)^{1/3}$, as the area LM of the rectangle $[0,L]\times[0,M]$ tends to infinity. Birgé has shown that $C_2/C_1 \leqslant 40$, which shows that the minimax risk is squeezed in rather tightly by the bounds in (3.17).

It was shown in Examples 2.2 and 3.2 that the minimax risk for the estimation of decreasing densities on [0,1], bounded by some $M > 0$ (which is the same for all densities in the class), tends to zero at the rate $n^{-1/3}$, as the sample size $n \to \infty$, if the loss is measured by L_1-distance. This suggests that a more precise picture of what is going on is obtained by looking at neighborhoods around a (decreasing) density f, which shrink at the rate $n^{-1/3}$, and by evaluating the (local) minimax risk of estimators based on a sample of size n over such a neighborhood. This leads to the following definition.

Definition 3.1.
Let $\mathcal{F}$ be a class of densities on $\mathbb{R}^d$ and let $E_f d_1(\hat{f}_n,f)$ be the risk under f of an estimator $\hat{f}_n$ of f based on a sample $X_1,\ldots,X_n$ from f, where d_1 denotes the L_1-distance. *Then the local asymptotic minimax risk* at a density $f \in \mathcal{F}$ is defined by

$$R_{LM}(f,d_1) = \sup_{c>0} \lim_{n\to\infty} \inf_{f_n} \sup_{g \in U_{n,c}(f)} n^{1/3} E_g d_1(\hat{f}_n,g) \tag{3.18}$$

where

$$U_{n,c}(f) = \{g \in \mathcal{F}: d_1(g,f) \leqslant c \cdot n^{-1/3}\}$$

We now have the following result.

THEOREM 3.1. *There exists a constant $c_1 > 0$ such that for each decreasing density f on [0,1], with a bounded continuous derivative f' such that $f' < 0$ on (0,1),*

$$R_{LM}(f,d_1) \geqslant c_1 \int_0^1 |f(t)f'(t)|^{1/3}dt, \tag{3.19}$$

where $R_{LM}(f,d_1)$ *is defined by* (3.18).

PROOF. We give the proof for the situation where $f' \leqslant a < 0$ and $f \geqslant b > 0$ on (0,1), but only minor changes are needed to give the proof for the situation where f' (or f) is allowed to tend to zero at the right endpoint of (0,1).

Let $x_0, x_1, \ldots, x_{2m}$ be an increasing sequence of points in [0,1] such that $x_0 = 0$, and

$$\begin{aligned} \delta_i &= x_{2i-1} - x_{2i-2} = x_{2i} - x_{2i-1} \\ &= \tfrac{1}{2} n^{-1/3} f(x_{2i-1})^{1/3} / |f'(x_{2i-1})|^{2/3} \end{aligned} \tag{3.20}$$

for $i = 1, \ldots, m$. Suppose that $[x_{2m-2}, x_{2m-1})$, $[x_{2m-1}, x_{2m})$ is the last pair of intervals of this type, contained in [0,1). *Although m,* δ_i*, and the points* $x_1, \ldots, x_{2m}$ *depend on n, we suppress this dependence to avoid cumbersome notation.* Furthermore, for ease of notation we put $y_i = x_{2i-1}$. Define the functions f_i and g_i on the interval $J_i = [y_i - \delta_i, y_i + \delta_i)$ by

$$f_i(x) = f(y_i), \quad x \in J_i, \tag{3.21}$$

and

$$g_i(x) = \begin{cases} f(y_i) + \delta_i |f'(y_i)|, & y_i - \delta_i \leqslant x < y_i \\ f(y_i) - \delta_i |f'(y_i)|, & y_i \leqslant x < y_i + \delta_i. \end{cases} \tag{3.22}$$

Let $\tilde{f}_n$ be a probability density on $[0, x_{2m})$ such that $\tilde{f}_n | J_i = k_n f_i$. Then $k_n \to 1$, as $n \to \infty$, implying that the function $\tilde{g}_n$, defined on $[0, x_{2m})$ by $\tilde{g}_n | J_i = k_n g_i$ will be non-negative and hence a probability density for n sufficiently large (since $\int_0^{x_{2m}} \tilde{g}_n(x)dx = \int_0^{x_{2m}} f_n(x)dx = 1$).

As $n \to \infty$, we have

$$\frac{1}{2} \int_{J_i} (\tilde{f}_n^{1/2} - \tilde{g}_n^{1/2})^2 \sim \frac{1}{4} \delta_i^3 f(y_i)^{-1} f'(y_i)^2 \tag{3.23}$$

$$\int_{J_i} |\tilde{f}_n - \tilde{g}_n| \sim 2\delta_i^2 |f'(y_i)| \tag{3.24}$$

Applying Assouad's Lemma we obtain by (3.20), (3.23) and (3.24)

$$\begin{aligned} R_{LM}(f,d_1) &\geqslant \lim_{n\to\infty} n^{1/3} \sum_i \delta_i^2 |f'(y_i)| \{1 - \sqrt{\tfrac{1}{2} n \delta_i^3 f'(y_i)^2 / f(y_i)}\} \\ &= \tfrac{3}{4} \lim_{n\to\infty} n^{1/3} \sum_i \delta_i^2 |f'(y_i)| \\ &= \tfrac{3}{8} \int_0^1 |f(x)f'(x)|^{1/3} dx. \quad \square \end{aligned}$$

The Grenander maximum likelihood estimator $\hat{f}_n$, to be discussed in section 4, has the property that for any "smooth" density f such that $f' < 0$ on (0,1),

$$\lim_{n\to\infty} n^{1/3} E_f \int |\hat{f}_n(t) - f(t)| dt = c \cdot \int_0^1 |f(t)f'(t)|^{1/3} dt, \tag{3.25}$$

where $c \approx 0.62$ (see GROENEBOOM (1984a). If $f(t) = 1$, $t \in [0,1]$ (the *uniform* density on [0,1]), the right-hand side of (3.25) is zero, and it can be shown that in this case

$$\lim_{n\to\infty} n^{1/2} E_f \int |\hat{f}_n(t) - f(t)| dt = \frac{\sqrt{\pi}}{2} \tag{3.26}$$

(see GROENEBOOM (1984a), Remark 3.2).

The behavior of kernel estimators is rather different. For example, it is proved in Devroye and Györfi that for *any* kernel estimator of the form (1.1) with a kernel K with bounded support

$$\hat{g}_n(t) = (nh)^{-1} \sum_{i=1}^{n} K((t-X_i)/h),$$

based on a sample $X_1,...,X_n$ generated by a density f, we have

$$\liminf_{n\to\infty} \inf_{h>0} n^{1/3} E_f \int |\hat{g}_n(t) - f(t)|dt \geqslant (8/(9\pi))^{1/3} \tag{3.27}$$

if f is the uniform density on [0,1] (Theorem 7, Ch. 5, DEVROYE and GYÖRFI (1985)). This shows that these kernel estimators can only achieve a rate of convergence $n^{-1/3}$, whereas the Grenander estimator achieves the rate $n^{-1/2}$.

More generally, it can be shown that the Grenander estimator achieves the rate $n^{-1/2}$ for any density f on [0,1], which only consists of flat parts and a finite number of jumps, whereas kernel estimators would only achieve rate $n^{-1/3}$ in this case.

A comparison of (3.25) and (3.19) indicates that the Grenander estimator has very good properties according to a (suitably defined) criterion of local minimax risk. However, at present it is still an unsolved problem how to choose the collection $\mathcal{F}$ of decreasing densities (and, for that matter, the corresponding neighborhoods $U_{n,c}(f)$ in (3.18)) in order to obtain nontrivial upper bounds for the local minimax risk. This has to do with the somewhat peculiar behavior of the functional $f \to \int_0^1 |f(t)f'(t)|^{1/3} dt$, and the fact that the convergence in (3.25) is certainly not uniform in f.

4. THE GRENANDER MAXIMUM LIKELIHOOD ESTIMATOR

DISTRIBUTION THEORY.

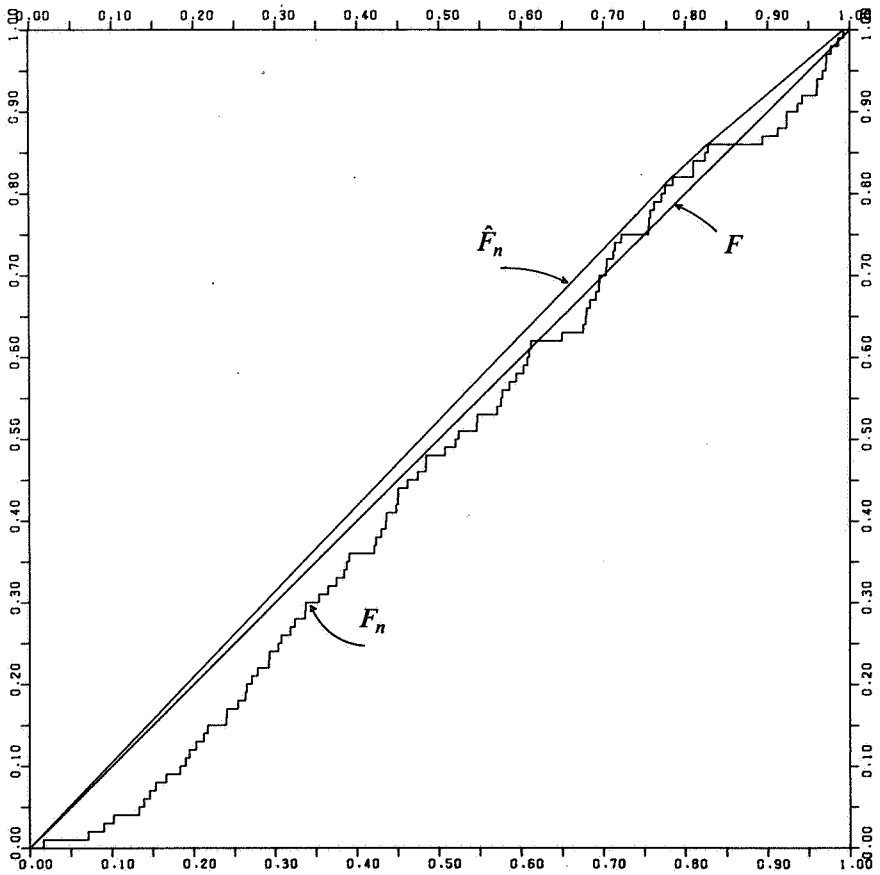
At the end of section 3 it was noticed that a particular density estimator "the Grenander maximum likelihood estimator" has a better performance in estimating decreasing densities than kernel estimators. We will now describe the construction of the Grenander estimator and we will offer an explanation for its good performance. The general consequences of an analysis of the behavior of the Grenander estimator are rather striking and not limited to the case of decreasing densities.

Suppose $X_1,...,X_n$ is a sample of n independent random variables generated by a density f on $[0,\infty)$. The *empirical distribution function* F_n of the sample is defined by

$$F_n(x) = \frac{1}{n} . \#\{i: X_i \leqslant x\}, \tag{4.1}$$

where $\#A$ denotes the number of elements in the set A. Thus $F_n(x)$ is the fraction of observations less than or equal to x. The *concave majorant* $\hat{F}_n$ of F_n on $[0,\infty)$ is by definition the smallest concave function $\geqslant F_n$ on $[0,\infty)$. Figure 4.1 shows a picture of the empirical distribution function F_n and its concave majorant $\hat{F}_n$ for a sample of $n=100$ observations, generated by the uniform density

$$f(t) = 1, \quad t \in [0,1]. \tag{4.2}$$



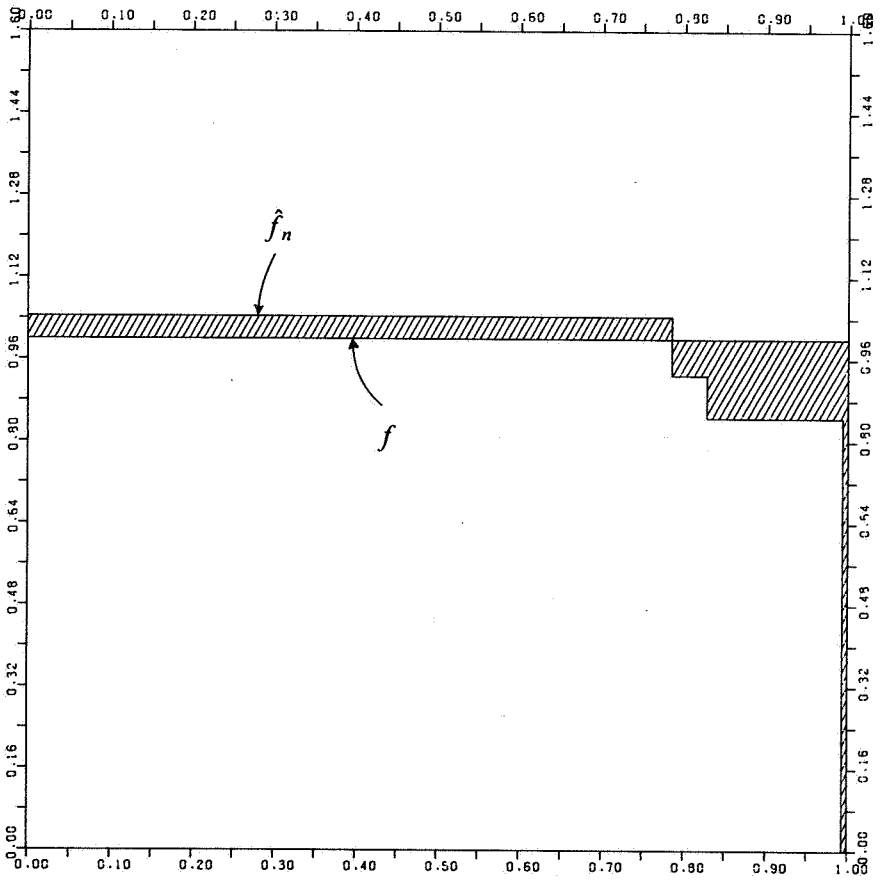
$n = 100.$ $F(t) = t,\ t \in [0,1].$

FIGURE 4.1. Concave majorant $\hat{F}_n$

Grenander shows (in GRENANDER (1956)) that the *maximum likelihood estimator* (MLE) of a decreasing density, based on a sample generated by this density, is given by the derivative of the concave majorant $\hat{F}_n$ of the empirical distribution F_n of the sample. Since the function $\hat{F}_n$ is piecewise linear with at most n changes of direction, the derivative is meaningful except at (at most) n points. Let $\hat{f}_n$ denote this derivative, defined at points of discontinuity by taking left-hand limits. This function satisfies

$$\prod_{i=1}^{n} \hat{f}_n(X_i) = \sup_{f \in \mathcal{F}} \prod_{i=1}^{n} f(X_i) \tag{4.3}$$

where $\mathcal{F}$ is the set of decreasing left-continuous densities on $[0, \infty)$. Thus $\hat{f}_n$ is *that* density f in the class $\mathcal{F}$ for which the "joint" density $\Pi_{i=1}^{n} f(X_i)$ at the observed points $X_1, \ldots, X_n$ attains its highest value, and for this reason $\hat{f}_n$ is called the maximum likelihood estimator. For a picture of $\hat{f}_n$, based on the same sample as used for figure 4.1, see figure 4.2.



$n=100, f(t)=1, t\in[0,1]$.

FIGURE 4.2. Grenander estimator.

The distribution theory of the Grenander estimator is still incomplete. An interesting early result is given in SPARRE ANDERSEN (1954), where it is proved that the number of jumps N_n of the function $\hat{f}_n$ is of order $\log n$, if the observations are generated by the uniform density defined by (4.2). More precisely, he proved that the distribution of the random variable $(N_n-\log n)/\sqrt{\log n}$ tends to a Gaussian distribution with mean zero and variance 1 (the "standard normal distribution"), as the sample size n tends to infinity. The proof in Sparre Andersen (1954) is based on rather elaborate enumeration techniques. At present it is possible to give a very quick proof of this result by using some properties of Brownian motion.

Since the further distribution theory of the Grenander estimator (and of density estimators in general) has been developed by using the relation between the empirical distribution function and Brownian motion, we now turn to an informal description of Brownian motion.

Let $X_1, X_2, \ldots$ be an infinite sequence of independent identically distributed random variables such that $P\{X_i=1\}=P\{X_i=-1\}=1/2$ for each i. For example, X_i could represent the outcome of the i-th trial in a fair coin-tossing game, where $X_i=1$ represents "heads" and $X_i=-1$ represents "tails". Corresponding to each infinite sequence $(X_1, X_2, \ldots)$ we define a function $W_n:[0,\infty)\to\mathbb{R}$ by

$$\begin{cases} W_n(0) = 0 \\ W_n(j/n) = n^{-1/2}\sum_{i=1}^{j} X_i, \quad j\in\mathbb{N} \end{cases} \tag{4.4}$$

and $W_n(t)$ is defined by linear interpolation for other values of $t \in [0,\infty)$. Each such function W_n is a possible realization of a *random walk* of a particle which jumps up or down according to the outcomes of a fair coin tossing game. By the central limit theorem we have that the distribution of $W_n(j/n)$ tends to a Gaussian distribution with mean zero and variance t, as $n\to\infty$ and $j/n\to t>0$. More generally, it has been shown by Wiener that one can define a limiting process consisting with probability one of continuous (nowhere differentiable) functions (or "paths") W on $[0,\infty)$ such that $W(t)$ has a Gaussian distribution with mean zero and variance t, for each t, and such that the distribution of $W(t)-W(s)$ is independent of that of $W(s)$ for $t>s$ (the process has *independent increments).* This process is called *Brownian motion* and can be considered as the limit (as $n\to\infty$) of the random walks W_n, defined by (4.4) on the basis of coin-tossing sequences $(X_1, X_2, \ldots)$.

The *Brownian bridge* on $[0,1]$ is a process of continuous paths $B:[0,1]\to\mathbb{R}$ which are obtained from Brownian motion paths W by the transformation

$$\begin{cases} B(t) = (1-t)W(t/(1-t)), & t\in[0,1) \\ B(1) \stackrel{def}{=} 0 \end{cases} \tag{4.5}$$

This transformation is called *Doob's transformation.* For a discussion of these concepts see e.g. BILLINGSLEY (1968), DOOB (1949) and ITÔ and MCKEAN (1974).

Brownian motion and the Brownian bridge arise in the context of density estimation in the following way. All the density estimators used in practice are based on the empirical distribution function F_n. Now it is already known for a long time (see e.g. DOOB (1949)) that the so-called *empirical process*

$$\sqrt{n}(F_n(t) - \int_0^t f(u)du), \quad t\in[0,\infty), \tag{4.6}$$

where F_n is the empirical distribution function based on a sample of size n generated by the density f on $[0,\infty)$, behaves for large n as a Brownian bridge with a changed time scale. More precisely it has been shown by KOMLÓS, MAJOR and TUSNÁDY (1975) that the supremum distance (over t) between the empirical process defined by (4.6) and a Brownian bridge process (with changed time scale)

$$\{B_n(F(t)), \quad t\in[0,\infty)\} \tag{4.7}$$

where $F(t) = \int_0^t f(u)du$, is smaller than $k \,.\, n^{-1/2}\log n$, with a probability tending to one as $n\to\infty$, for some fixed constant $k>0$. In particular we will have that well-behaving functionals of the empirical process will converge in distribution to the corresponding functional of the Brownian bridge; this is the so-called *invariance principle.* As an example, we have the following result.

THEOREM 4.1. *Let $\hat{f}_n$ be the Grenander density estimator, based on a sample of size n, generated by the uniform density f on [0,1] (see (4.2)). Then we have, as $n\to\infty$,*

$$n^{1/2}\int_0^1 |\hat{f}_n(t) - f(t)|dt \xrightarrow{d} 2\max_{t\in[0,1]} B(t), \tag{4.8}$$

i.e. the L_1-distance between $\hat{f}_n$ and f, multiplied by $n^{1/2}$, converges in distribution to 2 times the maximum of the Brownian bridge.

Sketch of proof.. Since $\hat{f}_n$ is the slope of the concave majorant $\hat{F}_n$ of the empirical distribution function F_n on $[0,1]$, we have that $n^{1/2}(\hat{f}_n - 1)$ is the slope of $n^{1/2}(\hat{F}_n - F)$, where $F(t) = \int_0^t f(u)du = \int_0^t du = t$, for $t\in[0,1]$.

This means that $S_n = n^{1/2}(\hat{f}_n - f)$ is the slope of the concave majorant of the empirical process $n^{1/2}(F_n - F)$ on $[0,1]$.

Applying the invariance principle, we get that the functional $\int_0^1 |S_n(t)|dt$ of the empirical process converges in distribution to the corresponding functional $\int_0^1 |S(t)|dt$ of the Brownian bridge, where $S(t)$ is the slope of the concave majorant of the Brownian bridge at t. But $\int_0^1 |S(t)|dt$ is just 2 times the maximum M of the Brownian bridge, since it is obtained by integrating $S(t)$ from 0 to the location ζ of the maximum and by integrating $-S(t)$ from ζ to 1. See Figure 4.3.

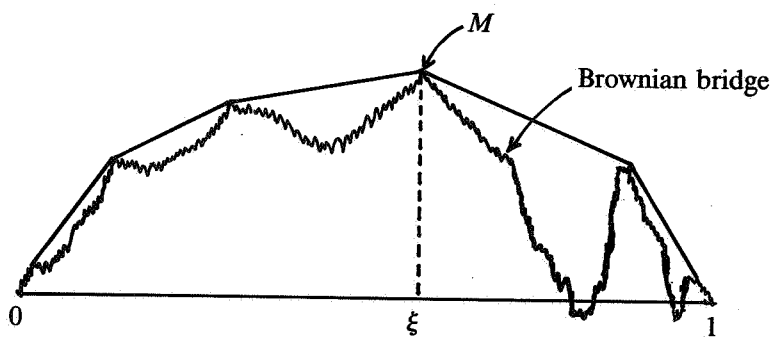


FIGURE 4.3.

(The slopes $S(t)$ will tend to $\infty(-\infty)$ as $t\downarrow 0$ ($t\uparrow 1$), which cannot be adequately shown in the picture.) □

Since the mean (or *first moment*) of the distribution of $2\max_{t\in[0,1]}B(t)$ equals $\sqrt{\pi/2}$, we obtain relation (3.26) as a corollary of Theorem 4.1. Similarly, by using the relation between the empirical process and the Brownian bridge, one can derive Sparre Andersen's result on the number of jumps of the Grenander estimator $\hat{f}_n$ if the underlying density is uniform (using the techniques of GROENEBOOM (1983)).

Theorem 4.1 is typical in the sense that the computation of the distribution of the functional $n^{1/2}\int_0^1|\hat{f}_n - f|dt$ of the empirical process $n^{1/2}(F_n - F)$ is transferred to the computation of the distribution of a corresponding functional of the Brownian bridge, but atypical in the sense that for functionals corresponding to density estimators we usually have to make a much closer (local) comparison between the behavior of the functionals for the empirical process and the Brownian bridge, using the results of KOMLÓS, MAJOR and TUSNÁDY (1975) (see the paragraph preceding Theorem 4.1). Also, the uniform density is a very "atypical" decreasing density, and the results are completely different if the density is strictly decreasing. In this case the "risk" $E_f\int|\hat{f}_n - f|dt$ decreases at a rate $n^{-1/3}$ (instead of $n^{-1/2}$). More precise information is given in the following theorem (Theorem 3.1 in GROENEBOOM (1984a)).

THEOREM 4.2. *Let f be a decreasing density, concentrated on a bounded interval $[0,B]$, with a bounded second derivative, and such that $f'(t)\neq 0$, for $t\in(0,B)$. Then there exists a constant $C=C(f)$ such that the distribution of the standardized L_1-distance*

$$n^{1/6}\{n^{1/3}\int_0^B|\hat{f}_n(t)-f(t)|dt - C\} \tag{4.9}$$

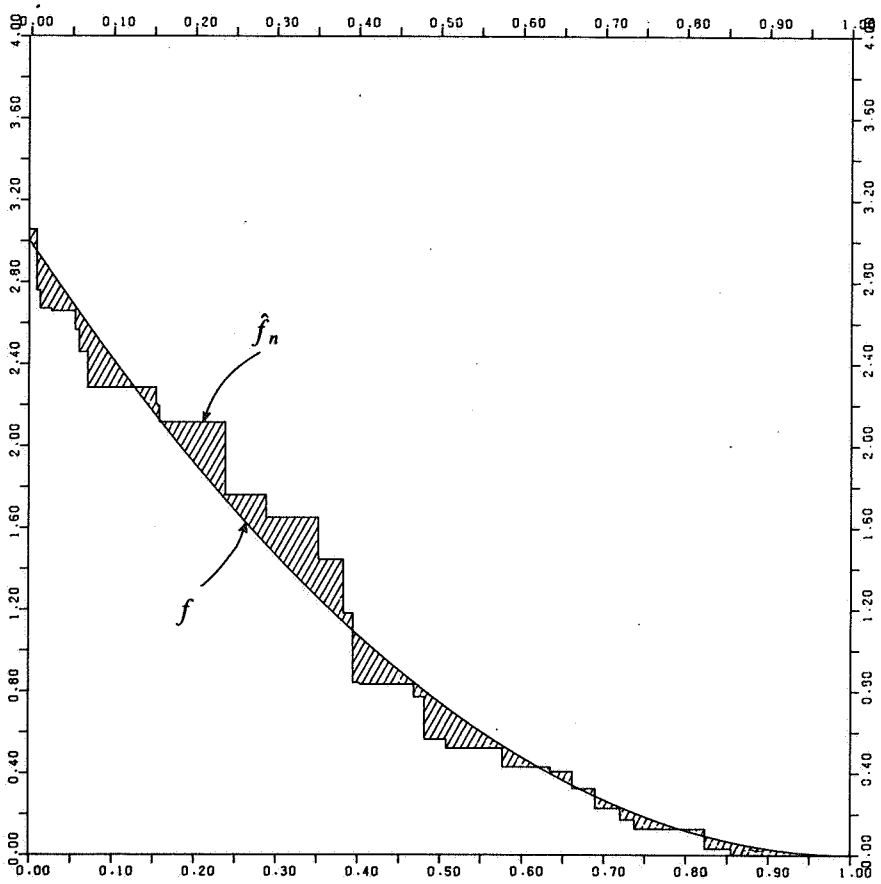
converges to a Gaussian distribution with mean zero.

The precise form of the constant C and the limiting Gaussian distribution cannot be given here, and the proof of this theorem is also omitted. However, we will try to describe informally the rather striking difference in behavior of the Grenander estimator $\hat{f}_n$ under the conditions of Theorem 4.1,

resp. Theorem 4.2. Figure 4.4 below shows a picture of the Grenander estimator $\hat{f}_n$ based on a sample of 1000 observation from the density f on [0,1], defined by

$$f(t) = 3(1-t)^2, \quad t \in [0,1], \tag{4.10}$$

which satisfies the conditions of Theorem 4.2.



$n = 1000$, $f(t) = 3(1-t)^2$, $t \in [0,1]$.

FIGURE 4.4. Grenander estimator.

In this case the number of jumps is of order $n^{1/3}$ (instead of order $\log n$, as in the case of the uniform density), and it can be shown that the number of jumps of $\hat{f}_n$ in an arbitrary interval $(c,d) \subset (0,1)$ will tend to infinity with probability one, as $n \to \infty$. In contrast to this, the number of jumps of $\hat{f}_n$ in each interval $(\epsilon, 1-\epsilon)$, $\epsilon > 0$, will remain *bounded* (in probability) as $n \to \infty$, if the underlying density f is uniform, and in this case the only cluster points will be 0 and 1 (for a picture, see figure 4.2). In the case of the density f, defined by (4.10), the curvature of the distribution function $F(t) = \int_0^t f(u)du, t \in [0,1]$, forces the concave majorant $\hat{F}_n$ of the empirical distribution function F_n to have many changes of direction, and as $n \to \infty$, the distributions of the derivatives $S_n(t)$ and $S_n(u)$ at two different points t and u of the interval (0,1) will become more and more independent. For the uniform density, there will be dependence over the whole interval, even as $n \to \infty$.

Thus the Grenander estimator "adapts" itself to the curvature of the underlying distribution whereas the usual kernel estimators don't have this property. This explains the better behavior of the Grenander estimator. Recently, there have been attempts to make the kernel estimators more "adaptive" (see HABBEMA et al. (1974), DUIN (1976), BREIMAN et al. (1977), CHOW et al. (1983), HALL

(1983)). For example, with the (kernel) density estimators proposed by HABBEMA et al. (1974) the window size is determined "adaptively", according to a criterion applied on the data set (the method of "cross-validation"). However, it seems clear that none of these adaptive kernel estimators can detect jumps of a density, whereas the Grenander estimator actually adapts itself both to jumps and to flat parts of a density. Also, the foregoing considerations apply to a much more general situation than the estimation of a monotone density, since, essentially, the discussed properties were based on *local* monotonicity of the density. So, although in the case of the estimation of non-monotone densities the Grenander estimator would no longer be applicable, we still are dealing *locally* with the random process on which the Grenander estimator in based *globally* in the case of a decreasing density. This process is a jump process of locations of maxima of Brownian motion with respect to a family of parabolas (the shape of which is determined by the underlying density; the structure of this process is determined in Section 4 of GROENEBOOM (1984b)). We will discuss the relevance of this process for the estimation of densities and distribution functions in a forthcoming paper.

ACKNOWLEDGMENT

I want to thank Lucien Birgé for inspiring conversations on the subject matter of the present paper.

REFERENCES

1 ASSOUAD, P., (1982). Classes de Vapnik-Çervonenkis et vitesse d'estimation. Preprint Université Paris à Orsay.

2 BEDNARSKI, T. (1982). Binary experiments, minimax tests and 2-alternating capacities, *Ann. Statist.* **10**, 226-232.

3 BILLINGSLEY, P., (1968). Weak convergence of probability measures, Wiley, New York.

4 BIRGÉ, L., (1980), Thèse, 3e partie. Université Paris VII.

5 BIRGÉ, L., (1983a). Approximation dans les espaces métriques et théorie de l'estimation, *Z. Wahrsch. Verw. Gebiete* **65**, 181-237.

6 BIRGÉ, L., (1983b). On estimating a density using Hellinger distance and some other strange facts. *MSRI-Preprint 45-83,* Berkeley.

7 BIRGÉ, L., (1983c). Estimating a density under order restrictions. Non-asymptotic minimax risk. *Preprint Université Paris X* - Nanterre.

8 BREIMAN, L., W. MEISEL & E. PURCELL, (1977). Variable kernel estimates of multivariate densities, *Technometrics* **19**, 119-137.

9 CHOQUET, G., (1953-1954), Theory of capacities, *Ann. Inst. Fourier* **5**, 131-292.

10 CHOQUET, G., (1959). Forme abstraite du théorème de capacibilité, *Ann. Inst. Fourier* **9**, 83-89.

12 CHOW, Y.S., GEMAN, S. & L.D. WU, (1983). Consistent cross-validated density estimation, *Ann. Statist.* **11**, 25-38.

13 DAVIS, P.J. & R. HERSH, (1981). The mathematical experience, Birkhäuser, Boston.

14 DEVROYE, L. & L. GYÖRfi, (1985). Nonparametric density estimation, the L_1 view, *to be published*, Wiley, New York.

15 DOOB, J.L., (1949). Heuristic approach to the Kolmogorov-Smirnov theorems, *Ann. Math. Statist.* **20**, 393-403.

16 DUIN, R.P.W., (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions, *I.E.E.E. Trans. Comput.* C-**25**, 1175-1179.

17 FELLER, W.F., (1968). An introduction to probability theory and its applications, *Vol. I* (3rd ed.), Wiley, New York.

18 GRENANDER, U., (1956). On the theory of mortality measurement, Part II, *Skand. Akt.* **39**, 125-153.

19 GROENEBOOM, P., (1983). The concave majorant of Brownian motion, *Ann. Probab.* **11**, 1016-1027.

20 GROENEBOOM, P. (1984a). Estimating a monotone density. Report MS-R8403, Centre for Mathematics and Computer Science, Amsterdam. *To appear in: Proceedings of the Neyman-Kiefer conference,* Berkeley, June-July 1983, Eds. Le Cam et al.

21 GROENEBOOM, P. (1984b). Brownian motion with a parabolic drift and Airy functions. Report MS-R8413, Centre for Mathematics and Computer Science, Amsterdam.

22 HABBEMA, J.D.F., J. HERMANS & K. VAN DEN BROEK, (1974). A stepwise discriminant analysis program using density estimation, In Compstat 1974, Ed. G. Bruckmann, pp. 101-110, Vienna, *Physica Verlag.*

23 HALL, P., (1983). Large sample optimality of least squares cross-validation in density estimation, *Ann. Statist.* **11**, 1156-1174.

24 HUBER, P.J. & V. STRASSEN, (1973), Minimax tests and the Neyman-Pearson lemma for capacities, *Ann. Statist.* **1**, 251-263.

25 IBRAGIMOV, I.A. & R.Z. HASMINSKII, (1980). Estimation of a distribution density (Russian), *Zap. Nauchn. Semin. LOMI* **98**, 61-85.
English translation in Journ. Sov. Math. **21**, (1983).

26 IBRAGIMOV, I.A. & R.Z. HASMINSKII, (1981). On nonparametric density estimates (Russian), *Zap. Nauchn. Semin. LOMI* **108**, 73-89. English translation to appear in Journ. Sov. Math.

27 IBRAGIMOV, I.A. & R.Z. HASMINSKII, (1981). Statistical estimation, asymptotic theory, *Springer*, Berlin.

28 ITÔ, K. & H.P. MCKEAN, JR, (1974). Diffusion processes and their sample paths., 2nd. ed., *Springer*, Berlin.

29 KOLMOGOROV, A.N. & V.M. TIKHOMIROV, (1961). ϵ-Entropy and ϵ-capacity of sets in function spaces, *Amer. Math. Soc. Transl. (2)* **17**, 277-364.

30 KOMLÓS, J., P. MAJOR & G. TUSNÁDY, (1975), An approximation of partial sums of independent r.v.'s and the sample d.f., *Z. Wahrsch. Verw. Gebiete* **32**, 111-131.

31 LE CAM, L., (1972). Limits of experiments, Proc. Sixth Berkeley Symp. on Math. Statist. and Probab. Vol. 1 Theory of statistics, 245-261, *Univ. of California press,* Berkeley.

32 LORENTZ, G.G. (1966). Metric entropy and approximation, *Bull. Amer. Math. Society* **72**, 903-937.

33 SPARRE ANDERSEN, E., (1954). On the fluctuation of sums of random variables II, *Math. Scand.* **2**, 195-223.