# 5

## DISTRIBUTION AND DENSITY ESTIMATION

**L. Devroye**
**McGill University, Montreal, Canada**

**L. Györfi**
**Budapest University of Technology and Economics, Budapest, Hungary**

## 5.1 Distribution estimation

The classical nonparametric example is the problem estimating a distribution function $F(x)$ from i.i.d. samples $X_1, X_2, \dots X_n$ taking values in $\mathcal{R}^d$ $(d \geq 1)$. Here on the one hand the construction of the empirical distribution function $F_n(x)$ is distribution-free, and on the other hand its uniform convergence, the Glivenko-Cantelli Theorem holds for all $F(x)$:

$$\lim_{n\to\infty} \sup_x |F_n(x) - F(x)| = 0 \text{ a.s.}$$

The Glivenko-Cantelli Theorem is really distribution-free, and the convergence in Kolmogorov-Smirnov distance means uniform convergence, so virtually it seems that there is no need to go further. However, if, for example, in a decision problem one wants to use empirical distribution functions for two unknown continuous distribution functions for creating a kind of likelihood then these estimates are useless. It turns out that we should look for stronger error criteria.

For this purpose it is obvious to consider the total variation: if $\mu$ and $\nu$ are probability measures on $\mathcal{R}^d$ then the total variation of $\mu$ and $\nu$ is defined by

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

where the supremum is taken over all Borel sets $A$.

However, if $\mu$ stands for the common distribution of $\{X_i\}$ and $\mu_n$ denotes the empirical distribution then for nonatomic $\mu$

$$V(\mu, \mu_n) = 1$$

a.s., so the empirical distribution is a bad estimate in total varition.

One may expect to find a more sophisticated sequence $\{\mu_n^*\}$ of distribution estimates of $\mu$ which is consistent in total variation:

$$\lim_{n\to\infty} V(\mu,\mu_n^*) = 0 \text{ a.s.}$$

**Theorem 5.1.** (DEVROYE AND GYÖRFI (1990)) *Given any sequence of distribution estimators $\{\mu_n^*\}$ there always exists a probability measure $\mu$ for which*

$$V(\mu,\mu_n^*) > 1/4 \textit{ for all } n \textit{ a.s.}$$

PROOF. This negative finding means that the total variation is a much stronger error criterion than the Kolmogorov-Smirnov distance such that it is impossible to construct a distribution estimate with distribution-free consistency in total variation. The proof borrows some arguments from Devroye (1983) and Rényi (1959). First, we need a rich family of singular continuous probability measures. The family of probability measures considered here is parametrized by a number $b \in [0,1]$ with binary expansion $b = 0.b_{(1)}b_{(2)}b_{(3)}\ldots$, $b_{(i)} \in \{0,1\}$. Let the binary random variables $Y_{(1)},Y_{(2)},\ldots$ be i.i.d. and uniformly distributed on $\{0,1\}$. We define the random variable $X = X(Y,b)$ by setting $X = 0.X_{(1)}X_{(2)}X_{(3)}\ldots$ in the ternary radix system used for $Y = 0.Y_{(1)}Y_{(2)}Y_{(3)}\ldots$, where

$$X_{(k)} = \begin{cases} 0, & \text{if } b_{(k)} = 0,\\ Y_{(k)}, & \text{if } b_{(k)} = 1.\end{cases}$$

Let $\mu_b$ denote the probability measure of $X = X(Y,b)$. If in the binary expansion of $b$ there are finitely many ($L$) zeros, then $\mu_b$ is absolutely continuous and distributes its mass uniformly on a set of Lebesgue measure $2^{-L}$.If in the binary expansion of $b$ there are finitely many ($L$) ones, then $\mu_b$ is discrete and puts its mass uniformly on a set of cardinality $2^L$. In other cases, $\mu_b$ is singular.

We write $X(Y_1,b),\ldots,X(Y_n,b)$ to denote a sample drawn from the distribution of $X(Y,b)$. We will replace $b$ at a crucial step in the argument by a uniform $[0,1]$ random variable $B$, which is independent of $Y_1,\ldots Y_n$. Put

$$A_k = \{0.x_{(1)}x_{(2)}\cdots : x_{(i)} \in \{0,1\} \text{ for all } i;\, x_{(k)} = 0\}.$$

Then

$$\mu_b(A_k) = \begin{cases} 1, & \text{if } b_{(k)} = 0,\\ 1/2, & \text{if } b_{(k)} = 1.\end{cases}$$

Let $\mu_n^*$ be an arbitrary distribution estimate based upon $X(Y_1,b),\ldots,X(Y_n,b)$. Let us now define the parameter estimate $b_n = 0.b_{n1}b_{n2}\ldots$ by its binary expansion with bits

$$b_{nk} = \begin{cases} 0, & \text{if } \mu_n^*(A_k) > 3/4,\\ 1, & \text{otherwise.}\end{cases}$$

Then

$$|\mu_n^*(A_k) - \mu_b(A_k)| \geq 1/4 I_{\{b_{nk} \neq b_{(k)}\}}.$$

Therefore

$$\begin{aligned}
\sup_b \inf_n V(\mu_n^*, \mu_b) &= \sup_b \inf_n \sup_A |\mu_n^*(A) - \mu_b(A)| \\
&\geq \sup_b \inf_n \sup_k |\mu_n^*(A_k) - \mu_b(A_k)| \\
&\geq \sup_b \inf_n \sup_k 1/4 I_{\{b_{nk} \neq b_{(k)}\}}.
\end{aligned}$$

Replace $b$ by $B$ and resulting $b_{nk}$ by $B_{nk}$. Then

$$\begin{aligned}
\sup_b \inf_n V(\mu_n^*, \mu_b) &\geq \inf_n \sup_k 1/4 I_{\{B_{nk} \neq B_{(k)}\}} \\
&= 1/4 \inf_n Z_n.
\end{aligned}$$

Our theorem is proved if we can show that $Z_n = 1$ almost surely for all $n$. Put $Z_{Nn} = I_{\{\cup_{k=1}^N [B_{nk} \neq B_{(k)}]\}}$. Then $Z_{Nn} \uparrow Z_n = I_{\{\cup_{k=1}^\infty [B_{nk} \neq B_{(k)}]\}}$. Therefore it suffices to show that

$$\lim_{N\to\infty} \mathbb{P}\{\cup_{k=1}^N [B_{nk} \neq B_{(k)}]\} = 1.$$

But $\mathbb{P}\{\cup_{k=1}^N [B_{nk} \neq B_{(k)}]\}$ is the error probability of the decision $(B_{n1}, \ldots, B_{nN})$ on $(B_{(1)}, \ldots B_{(N)})$ for the observations $X_1, \ldots, X_n$. For this decision problem the Bayes decision is

$$\tilde{B}_{nk} = \begin{cases} 0, & \text{if } X_{i(k)} = 0 \text{ for all } i = 1, \ldots, n, \\ 1, & \text{otherwise.} \end{cases}$$

Thus,

$$\begin{aligned}
\mathbb{P}\{Z_{Nn} = 1\} &= \mathbb{P}\{\cup_{k=1}^N [B_{nk} \neq B_{(k)}]\} \\
&\geq \mathbb{P}\{\cup_{k=1}^N [\tilde{B}_{nk} \neq B_{(k)}]\} \\
&= 1 - \left(1 - \frac{1}{2 \cdot 2^n}\right)^N \\
&\uparrow 1.
\end{aligned}$$

□

For distribution estimation we may consider other error criteria. Such error criteria can be derived from dissimilarity measures of probability measures, like $f$-divergences introduced by Csiszár (1967) (see also Liese, Vajda (1987) and Vajda (1989)). The three most important $f$-divergences in mathematical statistics are the total variation, the information divergence and the $\chi^2$-divergence.

If $\mu$ and $\nu$ are probability measures on $\mathcal{R}^d$ then the information divergence (or I-divergence, relative entropy, Kullback-Leibler number) of $\mu$ and $\nu$ is defined by

$$I(\mu,\nu) = \sup_{\{A_j\}} \sum_j \mu(A_j) \log \frac{\mu(A_j)}{\nu(A_j)},$$

where the supremum is taken over all finite Borel measurable partition $\{A_j\}$ of $\mathcal{R}^d$.

If $\mu$ and $\nu$ are discrete distributions then

$$I(\mu,\nu) = \sum_j \mu(\{j\}) \log \frac{\mu(\{j\})}{\nu(\{j\})}.$$

The following inequality, also called Pinsker's inequality, upperbounds the total variation in terms of I-divergence (cf. Csiszár (1967), Kemperman (1969) and Kullback (1967)):

$$2\{V(\mu,\nu)\}^2 \leq I(\mu,\nu) \tag{5.1}$$

If $\mu_n^* = \mu_n^*(.;X_1,\ldots,X_n)$ is a distribution estimate of $\mu$, then $\{\mu_n^*\}$ is said to be consistent in information divergence if

$$\lim_{n\to\infty} I(\mu,\mu_n^*) = 0 \; a.s.$$

By Pinsker's inequality (5.1), the information divergence dominates the total variation, so it follows from Theorem 5.1 that given any sequence of distribution estimators $\{\mu_n^*\}$ there always exists a probability measure $\mu$ for which the sequence $\{\mu_n^*\}$ is not consistent in information divergence. The situation is even worse, a discrete distribution with known support cannot be estimated consistently in information divergence:

**Theorem 5.2.** (GYÖRFI, PÁLI AND VAN DER MEULEN (1994)) *Assume that $\mu$ is a probability measure on the set of positive integers. Given any sequence of distribution estimators $\{\mu_n^*\}$ there always exists a probability measure $\mu$ with finite Shannon entropy*

$$H(\mu) = -\sum_{j=1}^{\infty} \mu(\{i\}) \log \mu(\{i\})$$

*for which*

$$I(\mu,\mu_n^*) = \infty \; a.s.$$

## 5.2 The density estimation problem

HOW TO MEASURE CLOSENESS. A random variable $X$ on $\mathcal{R}^d$ has a density $f$ if, for all Borel sets $A$ of $\mathcal{R}^d$, $\int_A f(x)\,dx = \mathbb{P}\{X \in A\}$. It thus serves as a tool for computing probabilities of sets. As it is a function that reveals the local concentration of probability mass, it may be used to visualize distributions of random variables. The statistician's problem, then, is to estimate $f$ from an i.i.d. sample $X_1, \ldots, X_n$ drawn from $f$. A density estimate is simply a mapping $f_n : \mathcal{R}^d \times (\mathcal{R}^d)^n \to \mathcal{R}^d$ (we write $f_n(x; X_1, \ldots, X_n)$ or $f_n(x)$). It is the global closeness of $f_n$ to $f$ that interests us. The choice of a density estimate is governed by a number of factors, like consistency, smoothness, ease of computation, interpretability, flexibility, robustness, versatility, and optimality for certain criteria. The early work in the field approached the problem largely as a functional estimation problem: $f$ was treated as any function, and tools from function approximation theory came to the rescue in the analysis of the performance of density estimates—Taylor series expansions played a key role, for example. The view we take in nonparametric density estimation is that $f$ is largely unknown and that no assumptions can be made about its properties. Methods or properties that are valid for all densities are said to be universal. It is quite surprising that there are density estimates that can approximate any density $f$ asymptotically in an appropriate sense. We see in section 5.3 that the histogram has this property. Other examples will follow.

The quality of a density estimate is measured by how well it performs the task at hand, estimating probabilities. In this respect, the total variation criterion is a natural distance:

$$\sup_{A \in \mathcal{B}} \left| \int_A f_n - \int_A f \right| ,$$

where $\mathcal{B}$ is the class of Borel sets of $\mathcal{R}^d$. If this is smaller than $\varepsilon$, then all probabilities will be estimated with errors not exceeding $\varepsilon$. We measure the closeness of two densities $f$ and $g$ by their $L_1$ distance $\int |f - g|$. There are many reasons for this, but all more or less follow from Scheffé's identity (Theorem 5.4 below): $\sup_{A \in \mathcal{B}} |\int_A f_n - \int_A f| = (1/2) \int |f - f_n|$. We can thus compare the performance of density estimates on an absolute scale, and $L_1$ distances indeed have a physical interpretation: if we know that $\int |f - g| < 0.04$, then we know that differences in probabilities are at most 0.02. In contrast, the interpretation of the inequalities involving other metrics such as the $L_2$ metric (example: $\int (f - g)^2 \leq 0.04$) or the Kullback-Leibler metric (example: $\int f \log(f/g) \leq 0.04$) in terms of probabilities is less obvious.

THE COUPLING DISTANCE. There is an interpretation of the $L_1$ distance in terms of samples that is interesting in its own right. Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be independent samples of identically distributed random variables, the former having density $f$, and the latter having density $g$. It should be clear that with probability one, all $2n$ are different. However, if we allow the $X$ and $Y$ samples to be dependent (while maintaining independence within each sample), it is possible to have $X_i = Y_i$ for many $i$'s. Let us introduce the Hamming distance

$$H_n = \sum_{i=1}^{n} \mathbb{I}_{[X_i \neq Y_i]} .$$

It is the number of $X_i$'s we need to change to turn the $X$-sample into the $Y$-sample. If $f = g$, then we can obviously take identical samples and achieve $H_n = 0$. Let $\mathcal{F}$ denote all possible joint distributions of $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ such that $X_1, \ldots, X_n$ is i.i.d. and drawn from $f$ and $Y_1, \ldots, Y_n$ is i.i.d. and drawn from $g$. $\mathcal{F}$ is called a coupling. Define the coupling distance by

$$C_n = \inf_{\mathcal{F}} \mathbb{E}\{H_n\} \, .$$

The coupling distance measures the minimal expected Hamming distance over all possible dependencies between samples. It is interesting to note that

$$C_n = \frac{n}{2} \int |f - g| \, .$$

Thus, $\int |f - g|/2$, a number between 0 and 1, is the fraction of a sample we need to alter in order to turn it in a sample with the other density. To see why this is true, consider the functions $m = \min(f, g)$, $(f-g)_+$ and $(g-f)_+$, and note that $f = m + (f-g)_+$, $g = m + (g-f)_+$. Also, $\int (f-g)_+ = \int (g-f)_+ = (1/2)\int |f-g|$, and $\int m = 1 - (1/2)\int |f-g|$. We can generate our two samples in stages. First we generate a binomial $(n, (1/2)\int |f-g|)$ random variable $N$, which decides what fraction of each sample is drawn from $(f-g)_+$ or $(g-f)_+$. The remainder, $n - N$, is drawn from $m$, and as it is common, we may make those values identical in both samples. If the samples are generated in this way, we have $H_n = N$, and $\mathbb{E}\{H_n\} = (n/2)\int |f-g|$. This shows that there is a coupling that achieves $C_n$. To see that we cannot improve over it, observe that $(f-g)_+$ and $(g-f)_+$ are functions of disjoint support.

INVARIANCE TO TRANSFORMATIONS. The total variation is invariant with respect to monotone transformations of the axes. In fact, let $T : \mathcal{R}^d \to S \subseteq \mathcal{R}^d$ be a bijection and a Borel-measurable mapping. Let $X$ and $Y$ be random variables with densities $f$ and $g$, respectively. Then the total variation distance is

$$\sup_B |\mathbb{P}\{X \in B\} - \mathbb{P}\{Y \in B\}|.$$

If $T$ is a bijection, this is equal to

$$\sup_B |\mathbb{P}\{T(X) \in T(B)\} - \mathbb{P}\{T(Y) \in T(B)\}|,$$

which is nothing but the total variation distance between $T(X)$ and $T(Y)$, provided that $\{T(B) : B \in \mathcal{B}\}$ coincides with the Borel sets of $S = T(\mathcal{R}^d)$. We leave this easy verification as an exercise.

The implications of this fact are tremendous. For example, if we know that a certain density estimate performs very well when the data belong to a compact set such as $[-1, 1]$, then we might apply the monotone mapping $T : x \to x/(1 + |x|)$, estimate the density of $T(X)$, and obtain an estimate of the density of $X$ by well-known back-transformation methods. Such estimates are called transformed density estimates. Here is another variation on the same theme: assume that we wish to visualize the densities $f$ and $g$ to get an idea of the size and the location of the error. Computer screens cannot show infinite tails, but by showing the graphs of the transformed densities (for a transformation to a compact set), we can make do, as the total variation distance remains unchanged.

THE LEBESGUE DENSITY THEOREM. Note that two densities $f$ and $g$ are identical if $\int_B f = \int_B g$ for all Borel sets $B$, that is, if and only if $\int |f-g| = 0$. Thus, we may alter $f$ on a set of zero Lebesgue measure, and still be in the same equivalence class of densities. This immediately makes criteria such as $|f(x)-g(x)|$ (pointwise error) or $\sup_x |f(x)-g(x)|$ (the supremum error) suspect, unless we replace these criteria by quantities that are defined in terms of these equivalence classes. Some help in this respect comes from the following theorem.

**Theorem 5.3.** (THE LEBESGUE DENSITY THEOREM) *Let $\mathcal{Q}$ be a subclass of the Borel sets of $\mathcal{R}^d$ with the property that*

$$\sup_{Q\in\mathcal{Q}} \frac{\lambda(Q^*)}{\lambda(Q)} \le c < \infty$$

*for some constant $c$, where $Q^*$ is the smallest cube centered at the origin that contains $Q$, and $\lambda(\cdot)$ denotes the volume of a set. (Here $a/0$ is treated as $\infty$ for all $a \ge 0$.) Let $\mathcal{Q}_r$ be the subclass of $\mathcal{Q}$ containing only sets $Q$ with $\lambda(Q) \le r$. Let $f$ be any density on $\mathcal{R}^d$. Let $z+Q$ denote the translation of $Q$ by $z$. Then for almost all $x$:*

$$\lim_{r\downarrow 0} \sup_{Q\in\mathcal{Q}_r} \left| \frac{1}{\lambda(Q)} \int_{x+Q} f(y)\,dy - f(x) \right| = 0.$$

*The points $x$ at which this convergence takes place are called Lebesgue points for $f$. Classes that satisfy the condition are the classes of all cubes, or all balls of $\mathcal{R}^d$.*

This theorem is typically applied to balls $B_{x,r}$ of radius $r$ centered at $x$. Let us define $g(x) = \liminf_{r\downarrow 0} \int_{B_{x,r}} f/\lambda(B_{x,r})$. The Lebesgue density theorem says that $g$ is in the equivalence class of $f$. And in fact, it is easy to see that all members of that equivalence class lead to the same $g$, and thus, $g$ may be considered as the representative of that class. We will call it the Lebesgue representative, denoted temporarily by $L(f)$. Now, of course, we may return to the pointwise criterion, and use $|L(f)(x)-L(g)(x)|$ instead of $|f(x)-g(x)|$. And we may use $\sup_x |L(f)(x)-L(g)(x)|$ as the new definition of supremum distance.

The Lebesgue density theorem states that for every density, without exception, for almost all $x$, $f(x)$ is close to an integral over a small ball centered at $x$. But integrals can be approximated by empirical measures based on data, and thus, the Lebesgue density theorem permits us to construct estimates that converge for all densities in the $L_1$ sense. Differentiation of measures, and approximation of functions by convolutions, is dealt with at length by Stein (1970), de Guzmán (1975, 1981), and Wheeden and Zygmund (1977).

Because of the following theorem, a distribution $\mu$ can be consistently estimated in total variation if it has a density.

**Theorem 5.4.** *(*SCHEFFÉ (1947)*) If $\mu$ and $\nu$ are absolutely continuous with respect to a $\sigma$-finite measure $\lambda$ with densities $f$ and $g$ respectively, then*

$$\|f-g\| := \int_{\mathcal{R}^d} |f(x)-g(x)|\lambda(dx) = 2V(\mu,\nu).$$

PROOF. Note that

$$\begin{aligned} V(\mu,\nu) &= \sup_A |\mu(A)-\nu(A)| \\ &= \sup_A |\int_A f d\lambda - \int_A g d\lambda| \\ &= \sup_A |\int_A (f-g) d\lambda| \\ &= \int_{f>g} (f-g) d\lambda \\ &= \int_{g>f} (g-f) d\lambda \\ &= \frac{1}{2}\int |f-g| d\lambda. \end{aligned}$$

□

The Scheffé Theorem results in a way of distribution estimation consistent in total variation via $L_1$-consistent density estimation: assume that $f_n$ is $L_1$-consistent, i.e.

$$\lim_{n\to\infty} \|f-f_n\| = 0 \text{ a.s.}$$

Introduce the distribution estimate induced by the density estimate $f_n$:

$$\mu_n^*(A) = \int_A f_n(x)\lambda(dx),$$

then

$$\lim_{n\to\infty} V(\mu,\mu_n^*) = 0 \text{ a.s.}$$

## 5.3 The histogram density estimate

The standard examples for $L_1$-consistent density estimates are the histogram and the kernel estimates.

Let $\mathcal{P}_n$ be a partition of $\mathcal{R}^d$ with cells $\{A_{n,j}\}$ of positive and finite Lebesgue measure. Then the **histogram** is as follows:

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

where $A_n(x) = A_{n,j}$ if $x \in A_{n,j}$.

**Theorem 5.5.** (ABOU-JAOUDE (1976)) *Assume that $\mu$ has a density $f$. If for each sphere $S$ centered at the origin*

$$\lim_{n\to\infty} \max_{j:A_{n,j}\cap S\neq\emptyset} diam(A_{n,j}) = 0 \tag{5.2}$$

*and*

$$\lim_{n\to\infty} \frac{|\{j : A_{n,j} \cap S \neq \emptyset\}|}{n} = 0 \tag{5.3}$$

*then*

$$\lim_{n\to\infty} \mathbb{E}\|f - f_n\| = 0$$

*and*

$$\lim_{n\to\infty} \|f - f_n\| = 0 \ a.s.$$

The beauty of Theorem 5.5 is that $L_1$-consistency is possible without any condition on the density $f$, thus we can have distribution estimate consistent in total variation if $\mu$ is absolutely continuous with respect to the Lebesgue measure. The histogram can be extended to distribution estimates consistent in total variation if the nonatomic part of $\mu$ is absolutely continuous with respect to a known $\sigma$-finite measure $\lambda$ (Barron, Györfi and van der Meulen (1992)).

The proof of Theorem 5.5 uses a denseness lemma:

**Lemma 5.1.** *The set of continuous functions of bounded support is dense in $L_1(\lambda)$.*

PROOF. We prove that for any $f \in L_1(\lambda)$ and $\varepsilon > 0$ there is a continuous function $g$ with compact support such that

$$\int |f(x) - g(x)|\lambda(dx) \leq \varepsilon.$$

Without loss of generality assume that $f \geq 0$. The case for arbitrary $f$ can be handled similarly as the case $f \geq 0$ by using the decomposition $f = f^+ - f^-$ where $f^+ = \max\{f, 0\}$ and $f^- = -\min\{f, 0\}$. $f \in L_1(\lambda)$ implies that there is a closed sphere $S = S_{0,R}$ centered at the origin with radius $R$ and $K > 0$ such that

$$\int |f(x) - \min\{f(x), K\}I_{x\in S}|\lambda(dx) \leq \varepsilon/2,$$

therefore it suffices to show that there is a continuous function $g$ with support in $S$ such that

$$\int_S |\min\{f(x), K\} - g(x)|\lambda(dx) \leq \varepsilon/2.$$

Put

$$f^*(x) = \min\{f(x), K\} I_{x \in S}$$

then using the technique of Lusin's Theorem we show that there is a continuous function $g$ with support in $S$ such that $0 \le g \le K$ and

$$\lambda\{x \in S; f^*(x) \ne g(x)\} \le \varepsilon/(2K).$$

Obviously this will imply the lemma since

$$\int_S |f^*(x) - g(x)| \lambda(dx) \le K\lambda\{x \in S; f^*(x) \ne g(x)\} \le K\varepsilon/(2K) = \varepsilon/2.$$

We have now that $0 \le f^* \le K$. Without loss of generality we may assume that $K = 1$. A function $s$ is called a *simple function* if its range consists of finitely many points in $[0, \infty)$. We can construct a monotonically increasing sequence of simple functions $s_1 \le s_2 \le \cdots \le f^*$ such that $s_n(x) \to f^*(x)$ as $n \to \infty$, for every $x \in S$. Indeed, for $n = 1, 2, \ldots$, and for $1 \le i \le 2^n$, define sets

$$E_{n,i} = \left\{ x \in S : \frac{i-1}{2^n} \le f^*(x) < \frac{i}{2^n} \right\}$$

and the simple function

$$s_n = \sum_{i=1}^{2^n} \frac{i-1}{2^n} I_{E_{n,i}}.$$

Sets $E_{n,i}$ are inverses images of half open intervals through a measurable function and thus are measurable. Clearly sequence $s_n$ is monotonically increasing and $s_n \le f^*$.

Now define $t_1 = s_1$ and $t_n = s_n - s_{n-1}, n = 2, 3, \ldots$. Then $2^n t_n$ is the indicator function of a set $T_n \subset S$ and

$$f^*(x) = \sum_{n=1}^{\infty} t_n(x).$$

Then there exist compact sets $U_n$ and open sets $V_n$ such that $U_n \subset T_n \subset V_n \subset S$ and $\lambda(V_n - U_n) < 2^{-n}\varepsilon/4$. Next we need a special version of Urysohn's Lemma which states that for any compact set $U$ and open set $V$ such that $U \subset V$ there exists a continuous function $h : S \to [0, 1]$ such that $h(x) = 1$ for $x \in U$ and $h(x) = 0$ for $x \in V^c$, where $V^c$ denotes the complement of $V$. In order to show this special case of Uryshon's Lemma for any set $A$ introduce the function

$$d(x, A) = \inf_{z \in A} \|x - z\|.$$

Then $d(x, A)$ is continuous, and $d(x, A) = 0$ iff $x$ lies in the closure of $A$. Such function $h$ can be defined as

$$h(x) = \frac{d(x, V^c)}{d(x, U) + d(x, V^c)}.$$

Thus by Uryshon's Lemma there are continuous functions $h_n$ that assume value 1 on $U_n$ and are zero outside $V_n$. Let

$$g(x) = \sum_{n=1}^{\infty} 2^{-n} h_n(x).$$

Since $h_n$ are bounded by 1 this series converges uniformly on $S$ and thus $g$ is continuous. Also its support lies in $S$. Since $2^{-n}h_n(x) = t_n(x)$ except in $V_n - U_n$, we obtain $g(x) = f^*(x)$ except in $\cup(V_n - U_n)$, but

$$\lambda(\cup(V_n - U_n)) \leq \sum_{n=1}^{\infty} \lambda(V_n - U_n) \leq \varepsilon/2 \sum_{n=1}^{\infty} 2^{-n} = \varepsilon/2.$$

□

PROOF OF THEOREM 5.5. By triangle inequality

$$\|f - f_n\| \leq \|f - \mathbb{E}f_n\| + \|\mathbb{E}f_n - f_n\|.$$

The first term of the right hand side is called bias, and the second one is the variation term.

$$\mathbb{E}f_n(x) = \mathbb{E}\frac{\mu_n(A_n(x))}{\lambda(A_n(x))} = \frac{\mu(A_n(x))}{\lambda(A_n(x))} = \frac{\int_{A_n(x)} f(z)\lambda(dz)}{\lambda(A_n(x))} =: T_n f(x).$$

If $f$ is uniformly continuous with bounded support then because of (5.2)

$$\|T_n f - f\| \to 0.$$

For arbitrary $f$, according to Lemma 5.1 choose continuous $\tilde{f}$ with bounded support such that

$$\|f - \tilde{f}\| < \varepsilon.$$

Then by the choice of $\tilde{f}$

$$\|T_n\tilde{f} - \tilde{f}\| \to 0.$$

Moreover

$$\begin{aligned}
\|T_n\tilde{f} - T_n f\| &= \int \left| \frac{\int_{A_n(x)} \tilde{f}(z)\lambda(dz)}{\lambda(A_n(x))} - \frac{\int_{A_n(x)} f(z)\lambda(dz)}{\lambda(A_n(x))} \right| \lambda(dx) \\
&= \sum_j \int_{A_{n,j}} \left| \frac{\int_{A_n(x)} \tilde{f}(z)\lambda(dz)}{\lambda(A_n(x))} - \frac{\int_{A_n(x)} f(z)\lambda(dz)}{\lambda(A_n(x))} \right| \lambda(dx) \\
&= \sum_j \left| \int_{A_{n,j}} \tilde{f}(z)\lambda(dz) - \int_{A_{n,j}} f(z)\lambda(dz) \right| \\
&\leq \sum_j \int_{A_{n,j}} |\tilde{f}(z) - f(z)|\lambda(dz) \\
&= \|\tilde{f} - f\|.
\end{aligned}$$

Thus

$$
\begin{aligned}
\|f-\mathbb{E}f_n\| &= \|f-T_n f\| \\
&\le \|f-\tilde{f}\|+\|T_n\tilde{f}-\tilde{f}\|+\|T_n\tilde{f}-T_n f\| \\
&\le 2\|f-\tilde{f}\|+\|T_n\tilde{f}-\tilde{f}\| \\
&\to 2\|f-\tilde{f}\| \\
&\le 2\varepsilon,
\end{aligned}
$$

therefore the bias term tends to zero. Observe that

$$
\begin{aligned}
\|\mathbb{E}f_n - f_n\| &= \int \left| \frac{\mu(A_n(x))}{\lambda(A_n(x))} - \frac{\mu_n(A_n(x))}{\lambda(A_n(x))} \right| \lambda(dx) \\
&= \sum_j |\mu(A_{n,j}) - \mu_n(A_{n,j})|.
\end{aligned}
$$

Fix a sphere $S$ centered at the origin such that $\mu(S^c) < \varepsilon$, and without loss of generality assume that $\{A_{n,j}\}$ are indexed such that $A_{n,j} \cap S \neq 0$ for $j = 1,2,\ldots,m_n$ and $A_{n,j} \cap S = 0$ otherwise. Put

$$
A_n = \cup_{j=m_n+1}^{\infty} A_{n,j}.
$$

By (5.3)

$$
\frac{m_n}{n} \to 0
$$

and $\mu(A_n) \le \mu(S^c) < \varepsilon$, thus

$$
\begin{aligned}
\|\mathbb{E}f_n - f_n\| &\le \sum_{j=1}^{m_n} |\mu(A_{n,j}) - \mu_n(A_{n,j})| + \mu(A_n) + \mu_n(A_n) \\
&\le \sum_{j=1}^{m_n} |\mu(A_{n,j}) - \mu_n(A_{n,j})| + |\mu(A_n) - \mu_n(A_n)| + 2\mu(A_n) \\
&\le \sum_{j=1}^{m_n} |\mu(A_{n,j}) - \mu_n(A_{n,j})| + |\mu(A_n) - \mu_n(A_n)| + 2\varepsilon.
\end{aligned}
$$

By the Cauchy-Schwarz and Jensen inequalities

$$
\begin{aligned}
\mathbb{E}\|\mathbb{E}f_n - f_n\| &\leq \sum_{j=1}^{m_n} \mathbb{E}|\mu(A_{n,j}) - \mu_n(A_{n,j})| + \mathbb{E}|\mu(A_n) - \mu_n(A_n)| + 2\varepsilon \\
&\leq \sum_{j=1}^{m_n} \sqrt{\mathbb{E}|\mu(A_{n,j}) - \mu_n(A_{n,j})|^2} + \sqrt{\mathbb{E}|\mu(A_n) - \mu_n(A_n)|^2} + 2\varepsilon \\
&\leq \sum_{j=1}^{m_n} \sqrt{\mu(A_{n,j})/n} + \sqrt{\mu(A_n)/n} + 2\varepsilon \\
&\leq \sqrt{(m_n+1)/n} + 2\varepsilon \\
&\to 2\varepsilon.
\end{aligned}
$$

Concerning the a.s. convergence, let $\mathcal{A}$ be the family of sets whose elements are the unions of $A_{n,1},\ldots,A_{n,m_n}$ and $A_n$ then by the Scheffé Theorem and Hoeffding's inequality

$$
\begin{aligned}
\mathbb{P}\{\|\mathbb{E}f_n - f_n\| > 3\varepsilon\} &\leq \mathbb{P}\left\{\sum_{j=1}^{m_n} |\mu(A_{n,j}) - \mu_n(A_{n,j})| + |\mu(A_n) - \mu_n(A_n)| > \varepsilon\right\} \\
&= \mathbb{P}\left\{2\sup_{A\in\mathcal{A}} |\mu(A) - \mu_n(A)| > \varepsilon\right\} \\
&\leq 2^{m_n+1} \sup_{A\in\mathcal{A}} \mathbb{P}\{|\mu(A) - \mu_n(A)| > \varepsilon/2\} \\
&\leq 2^{m_n+1} 2e^{-2n(\varepsilon/2)^2} \\
&= e^{-n(\varepsilon^2/2 - \ln 2\frac{m_n+2}{n})},
\end{aligned}
$$

which is summable, therefore applying the Borel-Cantelli Lemma the variation term tends to zero a.s. □

In the proof of Theorem 5.5, in fact, we have shown that

$$\mathbb{P}\{\|f - f_n\| > \varepsilon\} \leq 2e^{-n(\varepsilon^2/2 + o(1))}.$$

Using McDiarmid's inequality (Theorem 1.8) Devroye(1991) proved that for the histogram and the kernel density estimates (see Section 5.6)

$$\mathbb{P}\{|\|f - f_n\| - \mathbb{E}\|f - f_n\|| > \varepsilon\} \leq 2e^{-n\varepsilon^2/2}. \tag{5.4}$$

Using large deviation techniques the $L_1$ error of the histogram can be characterized as follows:

**Theorem 5.6.** (BEIRLANT, DEVROYE, GYÖRFI, VAJDA (2001)) *Assume (5.2). If there is a sequence of spheres $S_n$ centered at the origin such that $S_n \uparrow \mathcal{R}^d$ and*

$$\lim_{n\to\infty} \frac{|\{A_{n,j} \cap S_n \neq \emptyset\}|}{n} = 0 \tag{5.5}$$

*then for all* $0 < \varepsilon < 2$

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\{\|f - f_n\| > \varepsilon\} = -g(\varepsilon), \tag{5.6}$$

*where*

$$g(\varepsilon) = \min_{0<p<1-\varepsilon/2} \left( p \log \frac{p}{p+\varepsilon/2} + (1-p) \log \frac{1-p}{1-p-\varepsilon/2} \right).$$

Note that in Theorem 5.6 there is no condition on $f$, and so (5.6) holds for all $f$, and the rate function $g(\varepsilon)$ does not depend on $f$.

(5.4) implies that for consistent $f_n$

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\{\|f - f_n\| > \varepsilon\} \le -\varepsilon^2/2,$$

thus

$$g(\varepsilon) \ge \varepsilon^2/2.$$

We can get upper bound $\hat{g}(\varepsilon)$ on $g(\varepsilon)$ if in the definition of $g(\varepsilon)$ we substitute $p$ by $\frac{1-\varepsilon/2}{2}$. Then

$$\hat{g}(\varepsilon) = \frac{\varepsilon}{2} \log \frac{2+\varepsilon}{2-\varepsilon} \ge g(\varepsilon).$$

(Vajda (1970)). Further bounds can be found on p. 294-295 in Vajda (1989). For small $\varepsilon$ the upper and the lower bounds on $g(\varepsilon)$ are close to each other, and are approximately equal to $\varepsilon^2/2$.

In the proof of Theorem 5.6 we apply some lemmas, where we shall use the function

$$D(\alpha\|\beta) = \alpha \log \frac{\alpha}{\beta} + (1-\alpha) \log \frac{1-\alpha}{1-\beta}. \tag{5.7}$$

**Lemma 5.2.** (SANOV (1957), SEE P. 16 IN DEMBO, ZEITOUNI (1992), OR PROBLEM 1.2.11 IN CSISZÁR AND KÖRNER (1981)). *Let* $\Sigma = \{1, 2, \ldots, m\}$ *be a finite set (alphabet),* $\mathcal{L}_n$ *be a set of types (possible empirical distributions) on* $\Sigma$*, and let* $\Gamma$ *be a subset of* $\mathcal{L}_n$*. If* $Z_1, \ldots Z_n$ *are* $\Sigma$*-valued i.i.d. random variables with empirical distribution* $\mu_n$ *then*

$$\left| \frac{1}{n} \log \mathbb{P}\{\mu_n \in \Gamma\} + \min_{\tau\in\Gamma} I(\tau, \mu) \right| \le \frac{\log |\mathcal{L}_n|}{n} \tag{5.8}$$

*where* $|\mathcal{L}_n|$ *denotes the cardinality of* $\mathcal{L}_n$*.*

PROOF. We shall prove that

$$\mathbb{P}\{\mu_n \in \Gamma\} \le |\mathcal{L}_n| e^{-n \min_{\tau\in\Gamma} I(\tau,\mu)}$$

and

$$\mathbb{P}\{\mu_n \in \Gamma\} \ge \frac{1}{|\mathcal{L}_n|} e^{-n \min_{\tau\in\Gamma} I(\tau,\mu)}.$$

Because of our assumptions

$$\begin{aligned}
\mathbb{P}\{Z_1 = z_1, \ldots Z_n = z_n\} &= \prod_{i=1}^{n} \mathbb{P}\{Z_i = z_i\} \\
&= \prod_{i=1}^{n} \mu(z_i) \\
&= e^{\sum_{i=1}^{n} \log \mu(z_i)} \\
&= e^{\sum_{i=1}^{n} \sum_{j=1}^{m} I_{z_i=j} \log \mu(z_i)} \\
&= e^{\sum_{i=1}^{n} \sum_{j=1}^{m} I_{z_i=j} \log \mu(j)} \\
&= e^{\sum_{j=1}^{m} n\mu_n(j) \log \mu(j)} \\
&= e^{-n(H(\mu_n)+I(\mu_n,\mu))} \\
&=: \mathbb{P}_\mu(z_1^n).
\end{aligned}$$

For any probability distribution $\tau \in \mathcal{L}_n$ we can define a probability distribution $\mathbb{P}_\tau(z_1^n)$ in this way:

$$\mathbb{P}_\tau(z_1^n) := e^{-n(H(\mu_n)+I(\mu_n,\tau))}.$$

Put

$$T_n(\tau) = \{z_1^n : \mu_n(z_1^n) = \tau\},$$

then

$$1 \geq \mathbb{P}_\tau\{\mu_n = \tau\} = \mathbb{P}_\tau\{z_1^n \in T_n(\tau)\} = |T_n(\tau)| e^{-nH(\tau)}$$

therefore

$$|T_n(\tau)| \leq e^{nH(\tau)},$$

which implies the upper bound:

$$\begin{aligned}
\mathbb{P}\{\mu_n \in \Gamma\} &= \sum_{\tau \in \Gamma} \mathbb{P}_\mu\{\mu_n = \tau\} \\
&\leq |\mathcal{L}_n| \max_{\tau \in \Gamma} \mathbb{P}_\mu\{\mu_n = \tau\} \\
&= |\mathcal{L}_n| \max_{\tau \in \Gamma} |T_n(\tau)| e^{-n(H(\tau)+I(\tau,\mu))} \\
&\leq |\mathcal{L}_n| \max_{\tau \in \Gamma} e^{-nI(\tau,\mu)} \\
&= |\mathcal{L}_n| e^{-n \min_{\tau \in \Gamma} I(\tau,\mu)}.
\end{aligned}$$

Concerning the lower bound notice that for any probability distribution $\nu \in \mathcal{L}_n$

$$\begin{aligned}\frac{\mathbb{P}_\tau\{\mu_n=\tau\}}{\mathbb{P}_\tau\{\mu_n=\nu\}} &= \frac{|T_n(\tau)|\prod_{a\in\Sigma}\tau(a)^{n\tau(a)}}{|T_n(\nu)|\prod_{a\in\Sigma}\tau(a)^{n\nu(a)}}\\ &= \prod_{a\in\Sigma}\frac{(n\nu(a))!}{(n\tau(a))!}\tau(a)^{n(\tau(a)-\nu(a))}\\ &\geq 1.\end{aligned}$$

This last inequality can be seen as follows: the terms of the last product are of the forms $\frac{m!}{l!}\left(\frac{l}{n}\right)^{l-m}$. It is easy to check that $\frac{m!}{l!}\geq l^{m-l}$, therefore

$$\prod_{a\in\Sigma}\frac{(n\nu(a))!}{(n\tau(a))!}\tau(a)^{n(\tau(a)-\nu(a))}\geq\prod_{a\in\Sigma}n^{n(\tau(a)-\nu(a))}=n^{n(\sum_{a\in\Sigma}\tau(a)-\sum_{a\in\Sigma}\nu(a))}=1.$$

It implies that

$$\mathbb{P}_\tau\{\mu_n=\tau\}\geq\mathbb{P}_\tau\{\mu_n=\nu\}$$

and thus

$$1=\sum_\nu\mathbb{P}_\tau\{\mu_n=\nu\}\leq|\mathcal{L}_n|\mathbb{P}_\tau\{\mu_n=\tau\}=|\mathcal{L}_n||T_n(\tau)|e^{-nH(\tau)},$$

consequently

$$|T_n(\tau)|\geq\frac{1}{|\mathcal{L}_n|}e^{nH(\tau)}.$$

This implies the lower bound:

$$\begin{aligned}\mathbb{P}\{\mu_n\in\Gamma\} &= \sum_{\tau\in\Gamma}\mathbb{P}_\mu\{\mu_n=\tau\}\\ &\geq \max_{\tau\in\Gamma}\mathbb{P}_\mu\{\mu_n=\tau\}\\ &= \max_{\tau\in\Gamma}|T_n(\tau)|e^{-n(H(\tau)+I(\tau,\mu))}\\ &\geq \frac{1}{|\mathcal{L}_n|}\max_{\tau\in\Gamma}e^{-nI(\tau,\mu)}\\ &= \frac{1}{|\mathcal{L}_n|}e^{-n\min_{\tau\in\Gamma}I(\tau,\mu)}.\end{aligned}$$

**Lemma 5.3.** (BEIRLANT, DEVROYE, GYÖRFI, VAJDA (2001)) *Consider*

$$J_n=\sum_{j=1}^{m_n}|\mu(A_{n,j})-\mu_n(A_{n,j})|,$$

*based on a finite partition* $\mathcal{P}_n=\{A_{n,1},\ldots,A_{n,m_n}\}$*,* $(n\geq 2)$*, of* $\mathcal{R}^d$*. Assume*

$$\lim_{n\to\infty}m_n=\infty \tag{5.9}$$

*and*

$$\lim_{n\to\infty} \frac{m_n}{n} = 0 \tag{5.10}$$

*then for all* $0 < \varepsilon < 2$

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\{J_n > \varepsilon\} = -g(\varepsilon).$$

PROOF. We apply (5.8) for

$$\Sigma = \{A_{n,1}, \ldots, A_{n,m_n}\}$$

such that

$$\Gamma = \{\tau : 2V(\bar{\mu}_n, \tau) \geq \varepsilon\}.$$

Then, according to (5.8),

$$\left| \frac{1}{n} \log \mathbb{P}\{J_n \geq \varepsilon\} + \inf_{\tau\in\Gamma\cap\mathcal{L}_n} I(\tau, \bar{\mu}_n) \right| \leq \frac{\log |\mathcal{L}_n|}{n}.$$

Barron (1989) observed that

$$|\mathcal{L}_n| = \binom{n + m_n - 1}{m_n - 1} = e^{n\delta_n},$$

where $\delta_n \to 0$, provided (5.10). Thus, under (5.10),

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\{J_n > \varepsilon\} = -\lim_{n\to\infty} \inf_{\tau\in\Gamma\cap\mathcal{L}_n} I(\tau, \bar{\mu}_n).$$

It now remains to show that

$$\lim_{n\to\infty} \inf_{\tau\in\Gamma\cap\mathcal{L}_n} I(\tau, \bar{\mu}_n) = g(\varepsilon).$$

The distributions in $\mathcal{L}_n$ are possible empirical distributions, having components of the form $\frac{r}{n}$, where $r$ is integer. Because of (5.9) and because of the continuity of $V(\tau, \bar{\mu}_n)$ and $I(\tau, \bar{\mu}_n)$

$$\lim_{n\to\infty} \inf_{\tau\in\Gamma\cap\mathcal{L}_n} I(\tau, \bar{\mu}_n) = \lim_{n\to\infty} \inf_{2V(\tau,\bar{\mu}_n)\geq\varepsilon} I(\tau, \bar{\mu}_n).$$

Here

$$I(\tau, \bar{\mu}_n) = \sum_{j=1}^{m_n} \tau(A_{n,j}) \log \frac{\tau(A_{n,j})}{\mu(A_{n,j})}.$$

Put

$$L = \{j : \mu(A_{n,j}) > \tau(A_{n,j})\}$$

and

$$A_n = \cup_{j\in L} A_{n,j}.$$

Then

$$2V(\tau, \bar{\mu}_n) = 2(\mu(A_n) - \tau(A_n))$$

and, by the definition of I-divergence,

$$I(\tau,\bar{\mu}_n) \ge D(\tau(A_n)\|\mu(A_n)),$$

where the equality holds iff $\frac{\tau(A_{n,j})}{\mu(A_{n,j})}$ is constant both on $L$ and $L^c$. Thus

$$\begin{aligned}
\lim_{n\to\infty} \inf_{2V(\tau,\bar{\mu}_n)\ge\varepsilon} I(\tau,\bar{\mu}_n) &= \inf_{0<p<1-\varepsilon/2:\tau(A_n)=p,\mu(A_n)=p+\varepsilon/2} D(\tau(A_n)\|\mu(A_n)),\\
&= \inf_{0<p<1-\varepsilon/2}\left(p\log\frac{p}{p+\varepsilon/2}+(1-p)\log\frac{1-p}{1-p-\varepsilon/2}\right)\\
&= g(\varepsilon),
\end{aligned}$$

and Lemma 5.3 is proved.

PROOF OF THEOREM 5.6. Let $m_n$ and $A_n$ be as in the proof of Theorem 5.5. Put

$$\Delta_n = \sum_{j=1}^{m_n} |\mu(A_{n,j})-\mu_n(A_{n,j})| + |\mu(A_n)-\mu_n(A_n)|.$$

Then by Jensen's inequality

$$\begin{aligned}
\|f_n - f\| &= \sum_{j=1}^{\infty}\int_{A_{n,j}} |f_n(x)-f(x)|\lambda(dx)\\
&\ge \sum_{j=1}^{m_n}\int_{A_{n,j}} |f_n(x)-f(x)|\lambda(dx)\\
&+ |\sum_{j=m_n+1}^{\infty}\int_{A_{n,j}} f_n(x)\lambda(dx) - \sum_{j=m_n+1}^{\infty}\int_{A_{n,j}} f(x)\lambda(dx)|\\
&\ge \sum_{j=1}^{m_n}\left|\int_{A_{n,j}} f_n(x)\lambda(dx)-\int_{A_{n,j}} f(x)\lambda(dx)\right|\\
&+ \left|\int_{A_n} f_n(x)\lambda(dx)-\int_{A_n} f(x)\lambda(dx)\right|\\
&= \sum_{j=1}^{m_n} |\mu(A_{n,j})-\mu_n(A_{n,j})| + |\mu(A_n)-\mu_n(A_n)|\\
&= \Delta_n.
\end{aligned}$$

On the other hand

$$\begin{aligned}
\|f_n - f\| &\leq \|f_n - \mathbb{E}f_n\| + \|\mathbb{E}f_n - f\| \\
&= \sum_{j=1}^{\infty} |\mu(A_{n,j}) - \mu_n(A_{n,j})| + \|\mathbb{E}f_n - f\| \\
&\leq \sum_{j=1}^{m_n} |\mu(A_{n,j}) - \mu_n(A_{n,j})| + |\mu(A_n) - \mu_n(A_n)| \\
&\quad +2\mu(A_n) + \|\mathbb{E}f_n - f\| \\
&= \Delta_n + 2\mu(A_n) + \|\mathbb{E}f_n - f\|.
\end{aligned}$$

By definition $A_n \subset S_n^c$, and by assumption $S_n \uparrow \mathcal{R}^d$, therefore $\mu(A_n) \to 0$, so because of $\|\mathbb{E}f_n - f\| \to 0$

$$2\mu(A_n) + \|\mathbb{E}f_n - f\| \to 0,$$

and consequently Theorem 5.6 is proved if

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\{\Delta_n > \varepsilon\} = -g(\varepsilon), \tag{5.11}$$

which follows from Lemma 5.3 for $J_n = \Delta_n$.

For cubic partition assume that the density $f$ has a compact support and is continuously differentiable, then for consistent histogram $f_n$

$$\mathbb{E}(\|f_n - f\|) \leq \frac{c_1}{\sqrt{nh_n^d}} + c_2 h_n.$$

If $h_n = cn^{-1/(d+2)}$ then

$$\mathbb{E}(\|f_n - f\|) \leq Cn^{-1/(d+2)}.$$

(Devroye, Györfi (1985) and Beirlant, Györfi (1998)). For consistent histogram $f_n$

$$\frac{\|f_n - f\| - \mathbb{E}\|f_n - f\|}{\sqrt{Var(\|f_n - f\|)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

and

$$\limsup_{n\to\infty} \; nVar(\|f_n - f\|) \leq 1 - \frac{2}{\pi}$$

(Beirlant, Györfi and Lugosi (1994), Berlinet, Devroye and Györfi (1995) and Beirlant and Györfi (1998)).

If $\mu$ and $\nu$ are absolutely continuous with respect to a $\sigma$-finite measure $\lambda$ with densities $f$ and $g$ respectively, then the I-divergence of $\mu$ and $\nu$ becomes the so-called divergence between $f$ and $g$, i.e.

$$I(\mu,\nu) = \int_{\mathcal{R}^d} f(x) \log \frac{f(x)}{g(x)} \lambda(dx) = D(f,g).$$

Theorem 5.2 implies a negative result on density estimation for a dominated class:

**Theorem 5.7.** (GYÖRFI AND VAN DER MEULEN (1994)) *Given any sequence of density estimators $\{f_n\}$ there always exists a density $f$ with finite differential entropy*

$$H(f) = -\int f(x)\log f(x)dx$$

*and with arbitrary many derivatives such that*

$$D(f, f_n) = \infty \text{ a.s.}$$

Barron (1988) was one of the first to consider the problem of estimating a probability density function such that the density estimate is consistent in I-divergence. His results were generalized in Barron, Györfi and van der Meulen (1992) showing that if one imposes a certain condition on the class of distributions from which we are estimating the unknown one, namely that there exists a known probability measure $\nu$ such that $I(\mu,\nu) < \infty$, then one can construct a distribution estimator which is a.s. consistent in information divergence for all distributions in the class. As it is well-known, the condition $I(\mu,\nu) < \infty$ implies that $\mu$ is absolutely continuous with respect to $\nu$, so it is reasonable to create a distribution estimate, which has a density with respect to $\nu$. Introduce the notation

$$f(x) = \frac{d\mu}{d\nu}(x)$$

and for a partition $\mathcal{P}_n = \{A_{n,1}, A_{n,2}, ..., A_{n,m_n}\}$, $0 < m_n < n$, $n \geq 2$, assume that $\nu(A_{n,i}) \geq h_n = 1/m_n$. For a given sequence $\{a_n\}$, $0 < a_n < 1$ with

$$\lim_{n\to\infty} a_n = 0$$

consider the following density estimate:

$$f_n(x) = (1-a_n)\mu_n(A_n(x))/h_n + a_n, \tag{5.12}$$

For the choice

$$a_n = \frac{1}{nh_n + 1} \tag{5.13}$$

$f_n$ becomes a density estimator introduced by Barron (1988).

**Theorem 5.8.** (BARRON, GYÖRFI AND VAN DER MEULEN (1992)) *If $I(\mu,\nu) < \infty$ and*

$$\lim_{n\to\infty} h_n = 0, \quad \lim_{n\to\infty} nh_n = \infty.$$

*and*

$$\limsup \frac{1}{nh_na_n} \leq 1$$

*then*

$$\lim_{n\to\infty} D(f, f_n) = 0 \text{ a.s.}$$

The asymptotic normality of I-divergence of Barron's estimate has been proved by Berlinet, Györfi and van der Meulen (1997).

Berlinet (1995) made a comprehensive review on the asymptotic normality of the various global errors for nonparametric estimates.

## 5.4 Choosing Between Two Densities

Consider the following simple situation: $g_n$ and $f_n$ are two density estimates, and we must select the best one, that is, $\arg\min(\int |f_n - f|, \int |g_n - f|)$. More precisely, given the sample $X_1, \ldots, X_n$ distributed according to density $f$, we are asked to construct a density estimate $\phi_n$ such that

$$\int |\phi_n - f| \approx \min\left(\int |f_n - f|, \int |g_n - f|\right).$$

This simple problem turns out to be surprisingly difficult, even if the estimates $f_n$ and $g_n$ are fixed densities, not depending on the data.

The solution we propose here is fundamental, and is at the heart of the matter. First we introduce some notation. The empirical measure $\mu_n(A)$ of $\mu(A) \stackrel{\text{def}}{=} \int_A f$ is the measure that gives mass $\mu_n(A) = (1/n)\sum_{i=1}^n \mathbb{I}_{[X_i \in A]}$ to set $A$. The set $A = A(f_n, g_n) = \{x : f_n(x) > g_n(x)\}$ will be called the *Scheffé set* for the ordered pair $(f_n, g_n)$, as we recall Scheffé's identity,

$$\int |f_n - g_n| = 2\int (f_n - g_n)_+ = 2\int_A f_n - 2\int_A g_n,$$

valid whenever $\int f_n = \int g_n = 1$.

We define the *Scheffé estimate* $f_n^*$ as follows:

$$f_n^* = \begin{cases} f_n & \text{if } |\int_A f_n - \mu_n(A)| < |\int_A g_n - \mu_n(A)|, \\ g_n & \text{otherwise.} \end{cases}$$

Note that this is not a symmetric definition in $f_n$ and $g_n$! The situations in which we wish to carry out a selection are innumerable: we may want to pick the best of two bandwidths in kernel estimates, or the best of two summation sizes in series or wavelet estimates, or the best of a histogram and a kernel estimate, or the best of a parametric gamma density estimate and a non-parametric unimodal Grenander estimate. Scheffé's theorem was at the basis of our approach. The idea for selecting sets of the form $f > g$ when dealing with $L_1$ norms in density estimation was grabbed from Yatracos (1985) and developed by Devroye and Lugosi (1996, 1997) in the context of kernel density estimation.

**Theorem 5.9.** *Let $f_n$ and $g_n$ be two density estimates with $\int f_n = \int g_n = 1$. For the Scheffé* *estimate $f_n^*$, we have*

$$\int |f_n^* - f| \leq 3\min\left(\int |f_n - f|, \int |g_n - f|\right) + 4\max_{A\in\mathcal{A}}\left|\int_A f - \mu_n(A)\right|,$$

*where $\mathcal{A} = \{\{f_n > g_n\}, \{g_n > f_n\}\}$.*

PROOF. Denote by $\xi_n$ the best density estimate:

$$\xi_n = \begin{cases} f_n & \text{if } \int |f_n - f| < \int |g_n - f|, \\ g_n & \text{otherwise.} \end{cases}$$

Clearly,

$$\begin{aligned} \int |f_n^* - f| &\leq \int |\xi_n - f| + \int |\xi_n - f_n^*| \\ &= \int |\xi_n - f| + \int |f_n - g_n| \mathbb{I}_{[\xi_n \equiv f_n, f_n^* \equiv g_n]} \\ &\quad + \int |f_n - g_n| \mathbb{I}_{[\xi_n \equiv g_n, f_n^* \equiv f_n]} \\ &\stackrel{\text{def}}{=} I + II + III. \end{aligned}$$

We consider $II$. Let $E = [\xi_n \equiv f_n, f_n^* \equiv g_n]$ and $A = A(f_n, g_n)$:

$$\begin{aligned} II &= \int |f_n - g_n| \mathbb{I}_E \\ &= 2\left|\int_A f_n - \int_A g_n\right| \mathbb{I}_E \\ &= 2\left|\int_A f_n - \mu_n(A)\right| \mathbb{I}_E + 2\left|\int_A g_n - \mu_n(A)\right| \mathbb{I}_E \\ &\leq 4\left|\int_A f_n - \mu_n(A)\right| \mathbb{I}_E \\ &\leq 4\int_A |\xi_n - f| \mathbb{I}_{[E]} + 4\left|\int_A f - \mu_n(A)\right| \mathbb{I}_{[E]} \\ &\leq 2\int |\xi_n - f| + 4\left|\int_A f - \mu_n(A)\right|. \end{aligned}$$

For the middle step above, consider the cases $\int_A f_n > \mu_n(A) > \int_A g_n$ and $\int_A f_n > \int_A g_n > \mu_n(A)$ separately. Similarly, by switching $f_n$ and $g_n$, we see that

$$III \leq 2\int |\xi_n - f| + 4\left|\int_{A(g_n, f_n)} f - \mu_n(A(g_n, f_n))\right|.$$

Thus,

$$\int |f_n^* - f| \leq 3 \min\left(\int |f_n - f|, \int |g_n - f|\right) + 4 \sup_{A \in \mathcal{A}} \left| \int_A f - \mu_n(A) \right|.$$

$\square$

The Scheffé estimate thus has an error that is within

$$E_n \stackrel{\text{def}}{=} 4 \sup_{A \in \mathcal{A}} \left| \int_A f - \mu_n(A) \right|$$

of three times the best possible error. Note that this exceedance $E_n$ only depends upon $f_n$ and $g_n$ through the shape of the set $A$.

EXAMPLE: CHOOSING BETWEEN TWO NORMAL DENSITIES. Assume that $f_n$ and $g_n$ are both normal densities (perhaps determined by the data), and we must decide between them. Then $A$ is either of the form $(a,b)$ or $[a,b]^c$, and thus, $\mathcal{A}$ is contained in the collection $\mathcal{B}_2$ of unions of two intervals. But then,

$$\sup_{A \in \mathcal{B}_2} \left| \int_A f - \mu_n(A) \right| \leq 2 \sup_x |F_n(x) - F(x)|,$$

where $F$ is the distribution function for $F$, and $F_n$ is the empirical distribution function. But we know from Massart's tightening (1990) of the Dvoretzky–Kiefer–Wolfowitz inequality (1956) that

$$\mathbb{E}\left\{ \sup_x |F_n(x) - F(x)| \right\} \leq \frac{1}{\sqrt{n}}.$$

Therefore, if we use the empirical measure $\mu_n$, using *the same data* that were used in obtaining $f_n$ and $g_n$, regardless of the dependence of $f_n$ and $g_n$ upon these data, we have

$$\mathbb{E}\left\{ \int |f_n^* - f| - 3 \min\left(\int |f_n - f|, \int |g_n - f|\right) \right\} \leq \frac{8}{\sqrt{n}}$$

for all $n$. While the coefficient "3" may not be acceptable to some, we should stress nevertheless the simplicity of the procedure. Indeed, the set $A = \{f_n > g_n\}$ is very easily determined by the solution of a quadratic equation, and $\mu_n(A)$ is indeed easy to compute. Furthermore, the inequality above applies to any normal density estimate, with arbitrary dependence on the data, and to all densities.

Consider a more sophisticated situation, in which a normal density estimate $f_n$ is challenged by a histogram estimate $g_n$ having $k$ bins (of data-dependent widths, and data-dependent heights in the bins). Then both $A = A(f_n, g_n)$ and $A(g_n, f_n)$ consist of unions of at most $k+2$ disjoint intervals (we call the collection of such sets $\mathcal{B}_{k+2}$). Thus, by the argument presented above,

$$\sup_{A \in \mathcal{B}_{k+2}} \left| \int_A f - \mu_n(A) \right| \leq 2(k+2) \sup_x |F_n(x) - F(x)|.$$

Therefore, by Massart's inequality,

$$\begin{aligned}
\mathbb{E}\left\{\int |f_n^* - f| - 3\min\left(\int |f_n - f|, \int |g_n - f|\right)\right\} \\
&\le 8(k+2)\mathbb{E}\left\{\sup_x |F_n(x) - F(x)|\right\} \\
&\le \frac{8(k+2)}{\sqrt{n}}.
\end{aligned}$$

This inequality, while useful, is much too weak in general, when $k$ grows with $n$. This may be explained by noting its generality: it remains valid if a mean-spirited adversary is allowed to select the bins and bin heights in the worst possible manner so as to produce the worst possible selection $f_n^*$. If $\mu_n(A)$ in the definition of $f_n^*$ is based upon an independently drawn sample of size $m$, then in the two examples above, $\mathcal{A}$ would consist of two sets, and by the Cauchy–Schwarz inequality, for any fixed $A \in \mathcal{A}$:

$$\begin{aligned}
\mathbb{E}\left\{\left|\mu_n(A) - \int_A f\right| \Bigg| X_1, \ldots, X_n\right\} &\le \sqrt{\mathbb{E}\left\{\left(\mu_n(A) - \int_A f\right)^2 \Bigg| X_1, \ldots, X_n\right\}} \\
&= \mathbf{Var}\{\mu_n(A) | X_1, \ldots, X_n\} \le \sqrt{\frac{1}{4m}}.
\end{aligned}$$

Conditioning and then unconditioning, we have

$$\mathbb{E}\left\{\int |f_n^* - f| - 3\min\left(\int |f_n - f|, \int |g_n - f|\right)\right\} \le 4\sqrt{\frac{1}{m}}.$$

We must think of $m$ as a part of the data set aside for testing purposes, and will call these data test data. Typically, of course, $m$ is much smaller than $n$. In the former example (normal versus normal), this test data approach does not help. However, in the second example (normal versus any-histogram), the obtained bound is better than the previous one as soon as $m/n \ge 1/(k+2)^2$. So we distinguish between $\mu_n$ based on the original data and $\mu_n$ based on test data. Mostly, for simple parametric density estimates, the original data approach is best.

THE FACTOR THREE. One may wonder whether the factor three is necessary. It turns out that it cannot be removed as the following result from Devroye and Lugosi (2001) shows.

**Theorem 5.10.** *For any $t < 1$, $\varepsilon > 0$, there exist three densities, $g, h, f$, such that if $f_n = h$, $g_n = g$, then the Scheffé estimate $f_n^*$ based upon a sample of size $n$ drawn from $f$ has the following lower bound:*

$$\frac{\lim_{n\to\infty} \mathbb{E}\{\int |f_n^* - f|\}}{\min\left(\int |g - f|, \int |h - f|\right)} > 2 - \varepsilon.$$

*Furthermore,* $\lim_{n\to\infty} \mathbb{E}\int |f_n^* - f| = t$.

WHY NOT MAXIMUM LIKELIHOOD? In spite of the necessity of a constant factor larger than one, the Scheffé estimate is the only one we know for which the estimate is guaranteed to be within a constant factor of the smaller error plus a uniformly controllable term. One may wonder why we decided not to use another method of selection, such as the maximum likelihood. This is simply because the maximum likelihood criterion, even when applied under idealized circumstances, does not minimize $L_1$. Consider the real line, with two density estimates: $f_0 = \frac{1}{2}\mathbb{I}_{[-1,1]}$ and $g_\delta = \mathbb{I}_{[\delta,1+\delta]}$, where $\delta \in (0,1)$ is to be picked. Assume that the true density $f$ is the uniform density on $[0,1]$. If the data are $X_1,\ldots,X_n$, the maximum likelihood choice $\phi_n$ would be defined by

$$\phi_n \stackrel{\text{def}}{=} \arg\max\left(\prod_{i=1}^n f_0(X_i), \prod_{i=1}^n g_\delta(X_i)\right).$$

Clearly, everything depends upon $N$, the number of data points in $[0,\delta)$, a binomial $(n,\delta)$ random variable. We note that $g_\delta$ is picked if and only if $N=0$. Also, $\int|f_0-f|=1$, $\int|g_\delta-f|=2\delta$. Let $\delta < \frac{1}{2}$ so that $g_\delta$ is the better estimate. Clearly,

$$\begin{aligned}\mathbb{E}\left\{\int|\phi_n-f|\right\}-\int|g_\delta-f| &= (1-2\delta)\mathbb{P}\{N>0\}\\ &= (1-2\delta)(1-(1-\delta)^n)\\ &\to 1-2\delta,\end{aligned}$$

as $n\to\infty$. Thus, while for small $\delta$ the expected error for the best density is a respectable $2\delta$, the maximum likelihood estimate makes, with probability tending to one, a catastrophic blunder.

WHY NOT LEAST SQUARES? Others may want to minimize $\int(f_n-f)^2$, the $L_2$ distance. But even if we were given $f$, and even if that $f$ were square integrable, we might still make catastrophic choices. Consider this situation: $f$ is once again the uniform density on $[0,1]$, and we are indeed given this information. Let $\varepsilon$ be a small positive number, and assume that both density estimates ignore the data: $f_\varepsilon$ is our catastrophic candidate—it is $1-\varepsilon$ on $[0,1]$, and $\varepsilon^3$ on $[1,1+1/\varepsilon^2]$ (a long skinny tail, if you wish). Now, $g_\varepsilon = 1$ on $[-\varepsilon^2, 1-\varepsilon^2]$. Verify that $\int|g_\varepsilon - f| = 2\varepsilon^2 < 2\varepsilon = \int|f_\varepsilon - f|$, so that the choices are not even close! However, $\int(f_\varepsilon-f)^2 = \varepsilon^2+\varepsilon^4 < 2\varepsilon^2 = \int(g_\varepsilon-f)^2$, so that minimizing the $L_2$ error, even with $f$ given, picks the wrong density from the set $\{f_\varepsilon, g_\varepsilon\}$! Therefore, for universal properties in density estimation, criteria that are based on the square integrated distance are doomed. In contrast, the Scheffé estimate does not make such bad mistakes.

SELECTION FROM A FINITE SET. Consider now the selection problem with $k$ candidates $f_{ni}, 1\le i\le k$, such that $\int f_{ni}=1$ for all $i$. Let $A_{ij}, i<j$, denote the Scheffé set $A(f_{ni},f_{nj}) = \{x : f_{ni} > f_{nj}\}$.

We say that $f_{ni}$ wins against $f_{nj}$ when

$$\left|\int_{A_{ij}} f_{ni} - \mu_n(A_{ij})\right| < \left|\int_{A_{ij}} f_{nj} - \mu_n(A_{ij})\right|.$$

Of course, for each pair $i < j$, if $f_n^*$ denotes the winner, we have by Theorem 5.9:

$$\begin{aligned}
&\int |f_n^* - f| \\
&\qquad \leq 3\min\left(\int |f_{ni} - f|, \int |f_{nj} - f|\right) + 4 \sup_{A\in\{A_{ij},A_{ji}\}} \left|\int_A f - \mu_n(A)\right| \\
&\qquad = 3\min\left(\int |f_{ni} - f|, \int |f_{nj} - f|\right) + 4\Delta,
\end{aligned}$$

where

$$\Delta = \sup_{A\in\{A_{ij},A_{ji};1\leq i<j\leq k\}} \left|\int_A f - \mu_n(A)\right|.$$

Let us thus run a little competition with $k(k-1)/2$ matches, one for each ordered pair $i < j$. For each $f_{ni}$, we total the number of wins, and declare the density estimate with the maximum number of wins the Scheffé tournament winner. Ties are broken by picking the smallest index.

**Theorem 5.11.** *For the Scheffé tournament winner $f_n^*$, we have*

$$\int |f_n^* - f| \leq 9\min_i \int |f_{ni} - f| + 16\Delta.$$

PROOF. Let $m = \min_i \int |f_{ni} - f|$. Group the density estimates as follows: in group 0, we place all those with $\int |f_{ni} - f| = m$; in group 1, those with $\int |f_{ni} - f| \in (m, 3m + 4\Delta]$; in group 2, those with $\int |f_{ni} - f| \in (3m + 4\Delta, 9m + 16\Delta]$; in group 3, those with $\int |f_{ni} - f| > 9m + 16\Delta$. Let the number of density estimates in each group be denoted by $n_0, n_1, n_2$, and $n_3$, respectively. From Theorem 5.9, it is clear that any group 0 estimate must win against any group 2 or 3 estimate, and thus, their number of wins is at least $n_2 + n_3$. Any group 3 estimate must lose to any group 0 or 1 estimate, and thus, its number of wins is at most $n_2 + n_3 - 1$. Therefore, the Scheffé tournament winner cannot be from group 3. □

We may also declare another winner. We define

$$\Delta_i = \sup_{A\in\mathcal{A}} \left|\int_A f_{ni} - \mu_n(A)\right|,$$

where $\mathcal{A} = \{A_{ij}, A_{ji} : 1 \leq i < j \leq k\}$. The minimum distance estimate $\psi_n$ is that $f_{ni}$ of smallest index that minimizes $\Delta_i$. It is called a minimum distance estimate because it minimizes the distance to the empirical measure in a metric that is reminiscent of the total variation or $L_1$

distance—it would be the total variation distance if $\mathcal{A}$ were replaced by the class of Borel sets, $\mathcal{B}$. Only, the total variation distance

$$\sup_{B\in\mathcal{B}}\left|\int_B f-\mu_n(B)\right|=1$$

for any density $f$ and any empirical measure $\mu_n$: just let $B=\{X_1,\ldots,X_n\}$, for example. If each computation of $\int_A f_{ni}$ requires one time unit, then the computation of the Scheffé tournament winner requires $k(k-1)/2$ time units, while that for the minimum distance estimate requires $k^2(k-1)/2$ time units—it is about $k$ times computationally more intensive. This extra cost of course draws benefits, as is shown in the next theorem.

**Theorem 5.12.** *For the minimum distance estimate $\psi_n$, we have*

$$\int|\psi_n-f|\le 3\min_i\int|f_{ni}-f|+4\Delta.$$

PROOF. Let $\psi_n=f_{ni}$ and let $f_{nj}$ be any density estimate minimizing $\int|f_{n\ell}-f|$ over all $\ell$. Assume that $j\neq i$. Then, clearly,

$$\int|\psi_n-f|\le\int|f_{nj}-f|+\int|f_{ni}-f_{nj}|.$$

Now, assuming without loss of generality that $i<j$:

$$\begin{aligned}
\int|f_{ni}-f_{nj}| &= 2\sup_{A\in\{A_{ij},A_{ji}\}}\left|\int_A f_{ni}-\int_A f_{nj}\right|\\
&\le 2\sup_{A\in\mathcal{A}}\left|\int_A f_{ni}-\int_A f_{nj}\right|\\
&\le 2\sup_{A\in\mathcal{A}}\left|\int_A f_{ni}-\mu_n(A)\right|+2\sup_{A\in\mathcal{A}}\left|\int_A f_{nj}-\mu_n(A)\right|\\
&\le 4\sup_{A\in\mathcal{A}}\left|\int_A f_{nj}-\mu_n(A)\right| \quad \text{(by definition of } \psi_n=f_{ni}\text{)}\\
&\le 4\sup_{A\in\mathcal{A}}\left|\int_A f_{nj}-\int_A f\right|+4\sup_{A\in\mathcal{A}}\left|\int_A f-\mu_n(A)\right|\\
&\le 4\sup_{B\in\mathcal{B}}\left|\int_B f_{nj}-\int_B f\right|+4\Delta\\
&= 2\int\left|f_{nj}-f\right|+4\Delta
\end{aligned}$$

which together with the first inequality is all that is needed. □

SELECTING A NORMAL ESTIMATE. If we continue our examples, we note that if we are given any number $k$ of normal density estimates, then the class $\mathcal{A} \subseteq \mathcal{B}_2$, the class of intervals or complements of intervals. Thus, we have

$$\sup_{A\in\mathcal{B}_2}\left|\int_A f-\mu_n(A)\right| \le 2\sup_x |F_n(x)-F(x)|,$$

where $F$ is the distribution function for $F$, and $F_n$ is the empirical distribution function, we have, for $t>0$, if $\mu_n$ is the standard empirical measure,

$$\begin{aligned}\mathbb{E}\left\{\sup_{A\in\mathcal{A}}\left|\int_A f-\mu_n(A)\right|\right\} &\le \mathbb{E}\left\{\sup_{A\in\mathcal{B}_2}\left|\int_A f-\mu_n(A)\right|\right\}\\ &\le \mathbb{E}\left\{2\sup_x |F_n(x)-F(x)|\right\}\\ &\le \frac{2}{\sqrt{n}}.\end{aligned}$$

Thus, for the minimum distance estimate $\psi_n$:

$$\mathbb{E}\left\{\int|\psi_n-f|-3\min_i\int|f_{ni}-f|\right\}\le\frac{8}{\sqrt{n}}.$$

Note, in particular, that this bound does not depend upon $k$. In fact, we may select the best among $k$ arbitrary normal density estimates, and regardless of the value of $k$, we will obtain the same performance guarantees. In fact, Theorem 5.12 applies also to the infinite selection problem, although computations become a nightmare. With a selection from a possibly uncountably infinite class of densities, we can no longer index by integers $i$, and we can no longer be sure that minima are attained. The set-up must therefore be slightly modified. Also, we can no longer run Scheffé tournaments. Luckily, the minimum distance method remains valid.

## 5.5 The Minimum Distance Estimate

Let our density estimates be parametrized by $\theta \in \Theta$, where $\theta$ could represent the mean–variance pair in a normal density estimate, or the bandwidth in a kernel estimate, or the number of terms in a wavelet estimate. Let $f_{n,\theta}$ denote the density estimate with parameter $\theta$. Let

$$\mathcal{A}=\{\{f_{n,\theta}>f_{n,\theta'}\}:\theta\in\Theta,\theta'\in\Theta,\theta\ne\theta'\}.$$

Define

$$\Delta_\theta=\sup_{A\in\mathcal{A}}\left|\int_A f_{n,\theta}-\mu_n(A)\right|.$$

We define *the minimum distance estimate* $\psi_n$ as any density estimate selected from among those density estimates $f_{n,\theta}$ with

$$\Delta_\theta < \inf_{\theta^* \in \Theta} \Delta_{\theta^*} + 1/n.$$

The $1/n$ here is added to ensure the existence of such a density estimate.

**Theorem 5.13.** *Let $\{f_{n,\theta} : \theta \in \Theta\}$ be an arbitrary class of density estimates satisfying $\int f_{n,\theta} = 1$. For the minimum distance estimate $\psi_n$ as defined above for an infinite selection problem, we have*

$$\int |\psi_n - f| \le 3 \inf_{\theta \in \Theta} \int |f_{n,\theta} - f| + 4\Delta + \frac{3}{n},$$

*where $\Delta = \sup_{A \in \mathcal{A}} |\int_A f - \mu_n(A)|$.*

PROOF. Mimic the proof of Theorem 5.12. The details are left to the reader as an easy exercise. □

The theorem above makes a crucial connection between the error of the selected density estimate and the supremum error $\Delta$ for empirical probability measures over certain classes of sets. This permits us thus to use the rich theory of the uniform convergence of empirical measures initiated by Vapnik and Chervonenkis to density estimation.

Define $\mathcal{F} = \{f_n = f_\theta : \theta \in \Theta\}$. The Yatracos class $\mathcal{A}$ is the collection of all sets of the form $\{x : f_\theta(x) > f_{\theta'}(x)\}$, $\theta \ne \theta'$, and $\mu_n$ is the standard empirical measure based on the sample $X_1, \ldots, X_n$. Then Theorem 5.12 asserts that for all $f$:

$$\int |f_n - f| \le 3 \inf_{\theta \in \Theta} \int |f_\theta - f| + 4 \sup_{A \in \mathcal{A}} \left| \int_A f - \mu_n(A) \right| + \frac{3}{n}.$$

In particular, if $f \in \mathcal{F}$, then

$$\int |f_n - f| \le 4 \sup_{A \in \mathcal{A}} \left| \int_A f - \mu_n(A) \right| + \frac{3}{n}.$$

The right-hand side of the last inequality is of the order of $n^{-1/2}$ whenever $\mathcal{A}$ has a finite Vapnik–Chervonenkis dimension. The problem from this point on is purely combinatorial. In the rest of this section we work out examples for different classes.

SERIES ESTIMATES. Let $\psi_1, \ldots, \psi_k$ be fixed basis functions from $\mathcal{R}^d$ to $\mathcal{R}$ such that $\int \psi_i = r_i$ for all $1 \le i \le k$. We define the class $\mathcal{F}_k$ as the class of all linear combinations of the basis functions $f_\theta = \sum_{i=1}^k a_i \psi_i$ with coefficient vector $\theta = (a_1, \ldots, a_k)$ satisfying $\sum_{i=1}^k a_i r_i = 1$. The latter condition assures that all candidates $f_\theta$ have integral equal to one, a necessary condition to apply Theorem 5.12 for the performance of the minimum distance estimate. In this case, all sets $A_{\theta,\theta'} = \{x : f_\theta(x) > f_{\theta'}(x)\}$ are of the form

$$\left\{ x : \sum_{i=1}^{k} b_i \psi_i(x) > 0 \right\},$$

where $\theta' = (a'_1, \dots, a'_k)$, and $b_i = a_i - a'_i$. Lemma 1.6 shows that the Vapnik–Chervonenkis dimension of $\mathcal{A}$ is at most $k$, so invoking section 1.4.5, we obtain, for the minimum distance estimate $f_n$:

$$\mathbb{E}\int |f_n - f| \le 3 \inf_{g \in \mathcal{F}_k} \int |g - f| + c\sqrt{\frac{k}{n}} + \frac{3}{n},$$

where $c$ is a universal constant. For a survey on orthogonal series estimates and their $L_1$ aspects, we refer to Chapter 12 of Devroye and Györfi (1985). Surveys of approximation properties of orthogonal series may be found in Butzer and Nessel (1971).

PARAMETRIC ESTIMATES: EXPONENTIAL FAMILIES. Assume that the data are believed to be drawn from a normal density on the real line. This leads us to the construction of the class

$$\mathcal{F} = \left\{ f_{m,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2} : m \in \mathcal{R}, \sigma > 0 \right\}.$$

To determine the performance of the minimum distance estimate, we must estimate the Vapnik–Chervonenkis dimension of the corresponding Yatracos class $\mathcal{A}$. Note that every set in $\mathcal{A}$ may be written in the form of $\{x : ax^2 + bx + c \ge 0\}$ for some coefficients $a, b, c \in \mathcal{R}$, that is, $\mathcal{A}$ contains either closed intervals, or the union of two closed half-infinite intervals. The Vapnik–Chervonenkis dimension of this class is easily seen to be three.

The argument above may be significantly generalized as follows. A family $\mathcal{F}$ of densities on $\mathcal{R}^d$ is called an *exponential family* if each density in $\mathcal{F}$ may be written in the form

$$f_\theta(x) = c\alpha(\theta)\beta(x) e^{\sum_{i=1}^{k} \pi_i(\theta)\psi_i(x)},$$

where $\theta$ belongs to some parameter set $\Theta$, $\psi_1, \dots, \psi_k : \mathcal{R}^d \to \mathcal{R}$, $\beta : \mathcal{R}^d \to [0, \infty)$, $\alpha, \pi_1, \dots, \pi_k : \Theta \to \mathcal{R}$ are fixed functions, and $c$ is a normalizing constant. Examples of exponential families include classes of Gaussian, gamma, beta, Rayleigh, and Maxwell densities.

**Theorem 5.14.** *Let $\mathcal{F}$ be an exponential family of densities defined as above, and let $f_n$ be the minimum distance estimate based on this class. Then for any $f \in \mathcal{F}$:*

$$\mathbb{E}\int |f_n - f| \le c\sqrt{\frac{k+1}{n}} + \frac{3}{n},$$

*where $c$ is a universal constant. Moreover, for any density $f$:*

$$\mathbb{E}\int |f_n - f| \le 3 \inf_{\theta \in \Theta} \int |f_\theta - f| + c\sqrt{\frac{k+1}{n}} + \frac{3}{n}.$$

PROOF. Note that $f_\theta(x) > f_{\theta'}(x)$ if and only if

$$\sum_{i=1}^{k}(\pi_i(\theta)-\pi_i(\theta'))\psi_i(x)+\log\frac{\alpha(\theta)}{\alpha(\theta')}>0.$$

The class of functions appearing on the left-hand side spans a $(k+1)$-dimensional vector space, and so Lemma 1.6 implies that the Vapnik–Chervonenkis dimension of $\mathcal{A}$ is at most $k+1$. □

NEURAL NETWORK ESTIMATES. A feed-forward one-hidden-layer *neural network* (or simply neural network) with $k$ hidden nodes is a function $\phi : \mathcal{R}^d \to \mathcal{R}$ of the form

$$\phi(x)=\sum_{i=1}^{k}a_i\sigma(b_i^T x+c_i),$$

where $\sigma : \mathcal{R} \to \mathcal{R}$ is a fixed *activation function*, $a_i, c_i \in \mathcal{R}$, and $b_i \in \mathcal{R}^d$, $i = 1,\dots,k$, are the parameters of the network, and $b_i^T x$ denotes the usual inner product of the vectors $b_i$ and $x \in \mathcal{R}^d$. In order to generate a class eligible for the minimum distance method, we must require that $\int \sigma < \infty$. Define $\mathcal{F}$ as the set of neural networks $\phi$. Once again, to obtain a performance bound for the minimum distance density estimate, we need upper bounds for the Vapnik–Chervonenkis dimension of the associated Yatracos class. Each set in the class is of the form

$$\left\{x : \sum_{i=1}^{k}\left(a_i\sigma(b_i^T x+c_i)-a_i'\sigma(b_i'^T x+c_i')\right)>0\right\}.$$

Observe that $\mathcal{A}$ is a subclass of the class of all sets of the form $\{x : \psi(x) > 0\}$, where $\psi$ is a neural network of $2k$ hidden nodes. Computation of the Vapnik–Chervonenkis dimension of such classes of sets has been the subject of intensive research in pattern recognition and learning theory. It is known that the Vapnik–Chervonenkis dimension may be finite or infinite, depending on the activation function $\sigma$. The monograph of Anthony and Bartlett (1999) contains, among other things, the following useful result.

**Theorem 5.15.** (ANTHONY AND BARTLETT, 1999) *Let $\Theta \subset \mathcal{R}^m$ be a parameter set, and consider a class of functions $g_\theta : \mathcal{R}^d \to \mathcal{R}$ parametrized by $\Theta$. Assume that for each $\theta \in \Theta$ and $x \in \mathcal{R}^d$, $g_\theta(x)$ may be computed by at most $t$ steps of an algorithm which in each step executes one of the following operations: the arithmetic operations $+, -, \cdot$, and $/$ on real numbers; the exponential function $e^x$ on real numbers; an indicator function of the form $\mathbb{I}_{[x>a]}$, $\mathbb{I}_{[x\geq a]}$, $\mathbb{I}_{[x<a]}$, $\mathbb{I}_{[x\leq a]}$, $\mathbb{I}_{[x=a]}$, $\mathbb{I}_{[x\neq a]}$, for $a \in \mathcal{R}$. Then the Vapnik–Chervonenkis dimension $V$ of the class of sets $\{\{x : g_\theta > 0\} : \theta \in \Theta\}$ is at most*

$$V \leq t^2 m(m+19\log_2(9m)).$$

*If the exponential function is evaluated at most $q$ times for each pair $(x,\theta)$, then*

$$V \leq m^2(q+1)^2+11m(q+1)(t+\log_2(9m(q+1))).$$

The above theorem guarantees the finiteness of the Vapnik–Chervonenkis dimension of the neural network class for a whole host of all activation functions, though in many cases it does not provide the best possible bound. Theorem 5.15 may be used in many other situations, not just for neural network classes. Related results appear in Anthony and Bartlett (1999), who build on work of Goldberg and Jerrum (1995), Khovanskii (1991), Karpinsky and Macintyre (1997), Macintyre and Sontag (1993), and Koiran and Sontag (1997). Several related results are also surveyed in Devroye, Györfi, and Lugosi (1996).

MIXTURE CLASSES, RADIAL BASIS FUNCTION NETWORKS. Consider first the class of all mixtures of $k$ normal densities in $\mathcal{R}$, that is, the class of all densities of form

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{k} \frac{p_i}{\sigma_i} e^{-(x-m_i)^2/2\sigma_i^2},$$

where $(p_1, \ldots, p_k)$ is a probability vector, $\sigma_1, \ldots, \sigma_k$ are positive numbers, and $m_1, \ldots, m_k$ are arbitrary real numbers. The corresponding Yatracos class contains all sets of the form

$$\begin{aligned} &\{x : \phi(x) > \phi'(x)\} \\ &= \left\{ x : \sum_{i=1}^{k} \left( \frac{p_i}{\sigma_i} e^{-(x-m_i)^2/2\sigma_i^2} - \frac{p'_i}{\sigma'_i} e^{-(x-m'_i)^2/2\sigma_i'^2} \right) > 0 \right\}. \end{aligned}$$

This class of sets may be written in the form given in Theorem 5.15 with $m = 6k$, where the parameter set $\Theta$ is a subset of the $6k$-dimensional vector space of the parameters $p_i, p'_i, m_i, m'_i, \sigma_i, \sigma'_i$, $i = 1, \ldots, k$. Simple counting reveals that given $x$ and the vector of parameters, each function may be computed with no more than $t = 17k$ operations admitted by Theorem 5.15. Hence, we obtain that the Vapnik–Chervonenkis dimension of the Yatracos class based on all mixtures of $k$ univariate normal densities is bounded by a constant times $k^4$. We suspect that this estimate is loose. Nevertheless, the finiteness of the Vapnik–Chervonenkis dimension implies that the error of the minimum distance estimate is not more than a constant times $k^2/\sqrt{n}$, whenever the true density $f$ is, in fact, a mixture of $k$ normals.

The same argument may be generalized to the multivariate setting. Consider now the class $\mathcal{F}$ of all mixtures of $k$ normal densities over $\mathcal{R}^d$:

$$\phi(x) = \sum_{i=1}^{k} \frac{p_i}{\sqrt{(2\pi)^d \det(\Sigma_i)}} e^{-\frac{1}{2}(x-m_i)^T \Sigma_i^{-1}(x-m)},$$

where $(p_1, \ldots, p_k)$ is a probability vector, $\Sigma_1, \ldots, \Sigma_k$ are positive definite $d \times d$ matrices, and $m_1, \ldots, m_k$ are arbitrary elements of $\mathcal{R}^d$. Then it is clear from Theorem 5.15 that the Vapnik–Chervonenkis dimension of the corresponding Yatracos class is finite, and it is a matter of straightforward counting to obtain an explicit upper bound. In any case, if $f$ is a mixture of

$k$ normal densities, and is estimated by the minimum distance estimate based on the class defined here, then

$$\mathbb{E}\left\{\int |f_n - f|\right\} \le \sqrt{\frac{c_{k,d}}{n}},$$

where $c_{k,d}$ is a constant depending on $k$ and $d$. Radial basis function networks originated in the work of Poggio and Girosi (1990) and were analyzed in a series of subsequent papers, such as in Krzyżak and Linder (1998).

MIXTURES OF EXPONENTIALS. Often one may obtain much sharper estimates by direct methods. As an illustration, consider the class $\mathcal{F}$ of all mixtures of $k$ exponential densities (i.e., translations and scales of $e^{-x}, x \ge 0$). Again, it suffices to bound the Vapnik–Chervonenkis dimension of

$$\mathcal{A} = \{x : f(x) > g(x); f, g \in \mathcal{F}\}.$$

A member set in this class is thus of the form

$$\left\{x : \sum_{i=1}^{2k} a_i e^{-b_i x} I_{[x>c_i]} > 0\right\},$$

where $a_i, c_i \in \mathcal{R}$ and $b_i > 0$ are free parameters. Now clearly, we have at most $k+1$ intervals defined by the thresholds $c_i$. On the other hand, on each of these intervals, a set of the form $\{x : \sum_{i=1}^{2k} a_i e^{-b_i x} > 0\}$ defines at most $2k+1$ intervals (try showing this!). Therefore, each set in the class $\mathcal{A}$ is the union of at most $(2k+1)(k+1)$ intervals, and the Vapnik–Chervonenkis dimension of $\mathcal{A}$ is not more than $2(2k+1)(k+1) = O(k^2)$. Contrast this with the $O(k^4)$ bound obtainable by Theorem 5.15.

## 5.6 The Kernel Density Estimate

In this section, we get our first taste of real analysis, starting with some results on the approximations of functions in $L_1$. The problem is that $f$ cannot be approximated in $L_1$ by $\mu_n$, the empirical measure, as the total variation distance between any density $f$ and any atomic measure (like $\mu_n$) is 1. Thus, the approximation itself must have a density. The kernel estimate provides this: it smooths the empirical measure $\mu_n$.

We define the convolution density $f * K$ as the density of $X+Y$, where $X$ has density $f$ and $Y$ has density $K$ on $\mathcal{R}^d$. Think of this as a perturbation $Y$ applied to $X$. We note that

$$f * K(x) = \int f(z) K(x-z)\, dz.$$

The definition above also holds when $f$ and/or $K$ are absolutely integrable functions, and in that case we have

$$\int |f * K| \le \int |f| \times \int |K|,$$

a fact that is easy to check by change of integration, and will be referred to as Young's inequality. The convolution operation too lowers the total variation distance: for any densities $f, g$ and for any integrable function $K$:

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

To see this, we merely apply Young's inequality:

$$\int |f*K - g*K| = \int |(f-g)*K| \leq \int |f-g|*|K| = \int |f-g| \int |K|.$$

It is clear that if $Y$ (with density $K$) is concentrated near 0, then $f*K$ should be close to $f$. Indeed, we have the following fundamental approximation theorem from real analysis.

**Theorem 5.16.** *Let $K$ be an arbitrary integrable function on $\mathcal{R}^d$ (i.e., $\int |K| < \infty$), and let $f$ be a density on $\mathcal{R}^d$. Denoting $K_h(x) = (1/h^d)K(x/h)$, $x \in \mathcal{R}^d, h > 0$, we have*

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

PROOF. We may assume without loss of generality that $\int K \in \{0,1\}$. We will prove the statement when $\int K = 1$, leaving the $\int K = 0$ case as an easy exercise. Assume first that the statement is true for a dense subspace of functions $g$. Then

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

as $h \to 0$. Here we made use of Young's inequality:

$$\int |f*g| \leq \int |f| \cdot \int |g|$$

valid for any integrable functions $f, g$. The first term on the right-hand side can be made as small as desired by the choice of $g$ and the finiteness of $\int |K|$. So, we only need to prove the theorem for a dense subclass, such as the class of Lipschitz densities of compact support (cf. Lemma 5.1). Thus, let $f$ be Lipschitz with constant $C$ (i.e., $|f(x) - f(y)| \leq C\|x-y\|$, $x, y \in \mathcal{R}^d$), and supported on $[-M, M]^d$ for finite $M$. Let $L = KI_A$ where $I$ is the indicator function, and $A = [-r, r]^d$ is a large cube of our choice. Then, by Young's inequality,

$$\begin{aligned}
&\int |f * K_h - f| \\
&\quad \le \int \left| f * L_h - f \int L_h \right| + \int |f| \int |K_h - L_h| + \int |f * (K_h - L_h)| \\
&\quad \le \int \int |f(x-y) - f(x)| |L_h(y)| \, dy \, dx + 2 \int |K - L| \\
&\quad \le \int_{[-M-r, M+r]^d} \left( \int C \|y\| |L_h(y)| dy \right) dx + 2 \int |K - L| \\
&\quad \le (2M + 2r)^d Crh\sqrt{d} \int |L| + 2 \int |K - L| \\
&\quad \le (2M + 2r)^d Crh\sqrt{d} \int |K| + 2 \int |K - L| \\
&\quad = o(1) + 2 \int |K - L|,
\end{aligned}$$

which is as small as desired by choice of $r$ (and thus $L$). □

If our goal is to find a density estimate for which $\int |f_n - f|$ is small, we might make use of the fact that $\int |f * K_h - f|$ is small when $h$ is small and $\int K = 1$. Functions $K$ with $\int |K| < \infty$ and $\int K = 1$ will be called kernels. Indeed, $f * K_h$ may in turn be approximated by $\mu_n * K_h$, where $\mu_n$ is the empirical measure. More explicitly, $\mu_n * K_h$ is

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

which is nothing but the kernel estimate first proposed and developed by Akaike (1954), Parzen (1962), and Rosenblatt (1956). We may pick $h$ and $K$ as a function of $n$ and/or the data and write $h = h(X_1, \ldots, X_n)$, for example.

**Theorem 5.17.** (CONSISTENCY) *Let $K$ be a fixed kernel, and let $h$ depend on $n$ only. If $h \to 0$ and $nh^d \to \infty$ as $n \to \infty$, then $\mathbb{E}\{\int |f_n - f|\} \to 0$.*

PROOF. Note that

$$\int |f_n - f| \le \int |f * K_h - f| + \int |\mu_n * K_h - f * K_h|.$$

The first term on the right-hand side, also called the bias term, tends to 0 by Theorem 5.16. There are two tools that we will use repreatedly in the text. First of all, if $\nu$ is any probability measure, and $K$ and $L$ are kernels, then, by Young's inequality, $\int |\nu * K_h - \nu * L_h| \le \int |K - L|$. This inequality applies in particular when $\nu$ is $f$ and when it is $\mu_n$. Thus,

$$\int |\mu_n * K_h - f * K_h| \le \int |\mu_n * L_h - f * L_h| + 2 \int |K - L|.$$

The last term can be made as small as desired by choice of $L$ from among bounded kernels with support on a compact hypercube. The second trick uses the fact that $\int \nu * K_h = 1$ for any probability measure $\nu$ and kernel $K$. Thus,

$$\int |\mu_n * L_h - f * L_h| = 2\int (f * L_h - \mu_n * L_h)_+$$

and therefore

$$\begin{aligned}
&\mathbb{E}\left\{\int |\mu_n * L_h - f * L_h|\right\} \\
&= \int \mathbb{E}\{|\mu_n * L_h - f * L_h|\} \\
&= 2\int \mathbb{E}\{(f * L_h - \mu_n * L_h)_+\} \\
&= 2\int \min\left(f * L_h, \mathbb{E}\{(f * L_h - \mu_n * L_h)_+\}\right) \\
&\le 2\int \min\left(f * L_h, \sqrt{\mathbb{E}\{(f * L_h - \mu_n * L_h)^2\}}\right) \\
&\le 2\int \min\left(f, \sqrt{\mathbb{E}\{(f * L_h - \mu_n * L_h)^2\}}\right) + 2\int |f * L_h - f| \\
&= o(1).
\end{aligned}$$

The last step follows from $nh^d \to \infty$, the dominated convergence theorem, and Theorem 5.16. Here we used the fact that $\mu_n * L_h - f * L_h = \sum_{i=1}^n Z_i$, where the $Z_i$'s are i.i.d. zero mean random variables that have variance bounded by $(L_h)^2 * f/n^2$. Thus, $\mathbb{E}\{(\mu_n * L_h - f * L_h)^2\}$ is bounded by $(L^2)_h * f/(nh^d)$, which is $(f\int L^2 + o(1))/(nh^d)$ at all Lebesgue points for $x$. □

The remarkable thing about Theorem 5.17 is that it is valid whenever the kernel $\int K = 1$. For example, if we take $K$ to be the density that is uniformly distributed on the ball of radius 1 centered at $(1000, 0, \ldots, 0)$, then the average $f * K_h$ calculates an integral that does not even include the origin. Furthermore, it is not even necessary that $K$ be a density. Also, there is no possiblity of improving the conditions $h \to 0$ and $nh^d \to \infty$ (Devroye, 1983; Devroye and Györfi, 1985). In some cases, we have bandwidths that depend upon the data. The proof of the following theorem is left as an exercise.

**Theorem 5.18.** *Let $K$ be a fixed kernel. Let the bandwidth $H$ be an arbitrary function of the data such that $H \to 0$ and $nH^d \to \infty$ in probability as $n \to \infty$ (i.e., for every $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\{H + 1/(nH^d) > \varepsilon\} = 0$). Then $\mathbb{E}\{\int |f_n - f|\} \to 0$.*

Consider the quantity $\int |f_n - g|$ where $g$ is an arbitrary function of $x$ and may be substituted by $f$ or $f * K_h$. This quantity is very stable in the sense that if one $X_i$ changes value and the other

$n-1$ data points remain fixed (and denoting the new kernel estimate by $f_n^*$), then $\int |f_n - f_n^*| \le 2\int |K|/n$. By the bounded difference inequality (Theorem 1.8), we thus have

$$\mathbb{P}\left\{\left|\int |f_n - g| - \mathbb{E}\int |f_n - g|\right| \ge t\right\} \le 2e^{-nt^2/2(\int |K|)^2},\ t > 0.$$

In particular, we have

**Theorem 5.19.** (DEVROYE, 1987, 1988A, 1991) *If* $\sqrt{n}\mathbb{E}\int |f_n - f| \to \infty$, *then*

$$\frac{\int |f_n - f|}{\mathbb{E}\int |f_n - f|} \to 1 \text{ in probability.}$$

The concentration inequality in Theorem 5.19 shows that $\int |f_n - f| - \mathbb{E}\int |f_n - f|$ is of the order of $O(1/\sqrt{n})$. In fact, $\sqrt{n}(\int |f_n - f| - \mathbb{E}\int |f_n - f|)$ tends to a normal limit (Csörgő and Horváth, 1988; Beirlant and Mason, 1995).

Theorem 5.19 thus establishes the relative stability of the $L_1$ error for kernel estimates. It implies that the expected $L_1$ error is a good measuring stick because the actual error never deviates substantially from it. We will see further on that for all nonnegative kernels, the $\sqrt{n}$ condition is satisfied.

CHOOSING THE BANDWIDTH. In bandwidth selection, one is interested in functions $H = H(X_1, \ldots, X_n)$ such that $\int |f_{n,H} - f|$ comes close to $\inf_h \int |f_{n,h} - f|$, where $f_{n,h}$ makes the dependence upon $h$ explicit. A bandwidth $H$ with the property that

$$\limsup_{n\to\infty} \frac{\mathbb{E}\{\int |f_{n,H} - f|\}}{\inf_h \mathbb{E}\{\int |f_{n,h} - f|\}} \le C(f)$$

for some finite $C(f)$ is called a universal bandwidth. It is uniformly universal with constant $C$ if we may replace $C(f)$ by $C$ for all $f$. This problem has been attacked by many. However, many attempts can be discarded from the outset. Some authors suggested minimizing $\int (f_{n,h} - f)^2$ with respect to $h$. Even if $f$ were known, and even if $f$ were square integrable, the bandwidth thus selected would not be universally useful. For example, let us denote by $H$ a data-dependent bandwidth for which $\int (f_{n,H} - f)^2 \sim \inf_h \int (f_{n,h} - f)^2$ (so that $H$ is the $L_2$-optimal choice; for bounded densities, Stone (1984) shows how one can do this; Wegkamp (2000) provides a newer proof). Assume that $K$ is the uniform density on $[-1,1]$. For the Cauchy density $1/(\pi(1+x^2))$, $H \sim c/n^{1/5}$ in probability, yet one can check that $\mathbb{E}\{\int |f_{n,H} - f|\}/\inf_h \mathbb{E}\{\int |f_{n,h} - f|\} \to \infty$ as the optimal bandwidth for $L_1$ is larger than $n^{-1/5}$. In other words, even within the class of ultra-smooth densities (such as the Cauchy density), minimizing $L_2$ is just the wrong thing to do—intuitively, squaring tends to squash errors in the tails and make them unimportant.

Others have attempted to pick $h$ by maximum likelihood, for example by maximization of $\prod_{i=1}^n f_{n-1,i,h}(X_i)$, where $f_{n-1,i,h}$ is the kernel estimate based on the $n-1$ data points obtained after deleting $X_i$. While for densities with compact support on $\mathcal{R}$, the maximizing $H$ is indeed consistent, i.e., $H + 1/(nH^d) \to 0$ in probability (Devroye and Györfi, 1985), there is again no relationship with the total variation criterion. If the density $f$ has heavier than exponential tails, the method is not even consistent. The maximum likelihood cross-validation method was studied in detail by Broniatowski, Deheuvels and Devroye (1989).

Hall and Wand (1988) looked at the asymptotic expansion of $\mathbb{E}\{\int |f_{n,h} - f|\}$ and minimized the main asymptotic terms to obtain a recipe for $h$ as a function of $n$, $f$ and $K$. They then estimate the unknown quantity involving $f$ from the data, and propose this as a plug-in bandwidth estimate. For sufficiently smooth and small-tailed densities, and for positive kernels, they were able to show that $\mathbb{E}\{\int |f_{n,H} - f|\} \sim \inf_h \mathbb{E}\{\int |f_{n,h} - f|\}$. However, their method, and all other plug-in methods we are aware of, are not universal bandwidths.

The double kernel method uses a pair of kernels, $K$ and $L$, and picks $H = \arg\min_h \int |f_{n,h} - g_{n,h}|$, where $f_{n,h}$ and $g_{n,h}$ are the kernel estimates with kernels $K$ and $L$, respectively. Assume that $d = 1$. If the characteristic functions of $K$ and $L$ do not coincide on an open interval about the origin, then the choice $H$ is consistent (Devroye, 1989a). Furthermore, if $K$ and $L$ are symmetric, bounded kernels of compact support, $K \geq 0$, $\int x^2 L(x)\,dx = 0$, if both $K$ and $L$ are $L_1$-Lipschitz (i.e., $\int |K_1 - K_h| \leq C(h-1)$ for some $C < \infty$ and all $h > 1$, and similarly for $L$), and if $f$ is absolutely continuous, and $f'$ is absolutely continuous, $\int |f''| < \infty$ and $\int \sqrt{\sup_{|y|\leq 1} f(x+y)}\,dx < \infty$, then

$$\limsup_{n\to\infty} \frac{\mathbb{E}\{\int |f_{n,H} - f|\}}{\inf_h \mathbb{E}\{\int |f_{n,h} - f|\}} \leq \frac{1+\rho}{1-\rho}$$

where $\rho = 4\sqrt{\int L^2 / \int K^2}$ is a constant that can be made as small as desired by choice of $L$. The behavior outside the smoothness class described above is unknown: for example, we do not know whether the double kernel bandwidth is universal, let alone uniformly universal. The methodology studied in this lecture series and in Devroye and Lugosi (2001) is meant to fill the gap. It allows us to obtain a uniformly universal bandwidth with constant 3, and in fact, to obtain even stronger nonasymptotic uniformity:

$$\sup_f \frac{\mathbb{E}\{\int |f_{n,H} - f|\}}{\inf_h \mathbb{E}\{\int |f_{n,h} - f|\}} \leq 3 + o(1).$$

Note in particular that this result does not depend on the dimension, and that the bound is uniform over all $f$. Bandwidth selection methods are surveyed by Berlinet and Devroye (1994), Cao, Cuevas and González-Manteiga (1994), and Devroye (1997). A detailed study of the $L_1$ behavior of the kernel estimate in $\mathcal{R}^d$ is provided by Holmström and Klemelä (1992).

CHOICE OF THE KERNEL. The selection of the pair $(h, K)$ may be tackled as one problem, and indeed, $h$ may be absorbed into $K$ as a scale factor. Asymptotic evidence for smooth densities shows that for large sample sizes, the shape of the optimal kernel is unique. For example, for

$\mathcal{R}^1$, classical $L_2$ theory (Watson and Leadbetter, 1963) shows that for $L_2$ errors, among all positive kernels, the Epanechnikov kernel (Epanechnikov, 1969) $K(x) = \max(3/4(1-x^2), 0)$ is best possible. For $\mathcal{R}^d$, Deheuvels (1977) showed the $L_2$ optimality of $c\max((1-\|x\|^2)^d, 1)$. For the $L_1$ error, there is evidence that the Epanechnikov kernel is also best among all positive kernels (see the discussion in the next section). For these reasons, authors have typically decoupled the choices of $K$ and $h$, and picked $K$ fixed as one of these asymptotically optimal kernels.

RATES OF CONVERGENCE. The rate of convergence of the $L_1$ error of the standard kernel estimate is well understood in $\mathcal{R}$ but much less so in $\mathcal{R}^d$. For simplicity, we stick to $\mathcal{R}$ in this section. In that case, we note the following:

A. $\inf_f \liminf_n \inf_h n^{2/5}\mathbb{E}\{\int |f_{nh} - f|\} \geq A(K)$ where $A(K) \geq 0.86$ is a function of $K$ only, and is minimized for the Epanechnikov kernel $K(x) = 3/4(1-x^2)_+$ (Devroye and Penrod, 1984; see also Devroye and Györfi, 1985). The $n^{-2/5}$ rate is thus a universal lower bound beyond which we cannot go within the class of standard kernel densities.

B. There is no uniform rate of convergence: $\sup_f \inf_h \mathbb{E}\{\int |f_{nh} - f|\} = 2$.

C. There is no universal rate of convergence for individual densities: for any sequence $a_n \downarrow 0$, there exists a density $f$ such that for all $n$ large enough,

$$\inf_h \mathbb{E}\int |f_{nh} - f| > a_n.$$

See Birgé (1986), Devroye (1983, 1995).

## 5.7 Additive Estimates and Data Splitting

Assume that we are given a class of density estimates parametrized by $\theta \in \Theta$, such that $f_{n,\theta}$ denotes the density estimate with parameter $\theta$. Our goal is to construct a density estimate $f_n$ whose $L_1$ error is (almost) as small as that of the best estimate among the $f_{n,\theta}$, $\theta \in \Theta$. Applying the minimum distance estimate directly to this class is often problematic because of the dependence of each estimate in the class and the empirical measure $\mu_n$. Consider, as a basic example, the class of kernel estimates

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

parametrized by the smoothing factor $h \in (0,\infty)$, where $K$ is a fixed nonnegative function with $\int K = 1$. It is easy to see that if

$$\mathcal{A}_\Theta = \mathcal{A} = \{\{f_{n,h} > f_{n,h'}\} : h, h' \in (0,\infty), h \neq h'\},$$

then

$$\sup_{A\in\mathcal{A}}\left|\int_A f_{n,h}-\mu_n(A)\right|=1$$

for all $h$, so the minimum distance method selects a degenarate estimate.

We remedy this problem by introducing artificial independence between the estimates and the empirical measure. This may be achieved by holding out $m$ samples from the design of the density estimates, and using the empirical measure based on the held-out samples to construct the minimum distance estimate. More precisely, let $m<n$, and define $\mathcal{A}_\Theta$ as the *Yatracos class* of subsets of $\mathcal{R}^d$ (corresponding to the family of density estimates $f_{n,\theta}$, $\theta\in\Theta$) as the class of all sets of the form

$$A_{\theta_1,\theta_2}=\left\{x: f_{n-m,\theta_1}(x)>f_{n-m,\theta_2}(x)\right\},\ \theta_1,\theta_2\in\Theta.$$

We select a parameter $\theta_n$ from $\Theta$ by minimizing the distance

$$\Delta_\theta=\sup_{A\in\mathcal{A}_\Theta}\left|\int_A f_{n-m,\theta}-\mu_m(A)\right|.$$

over all $\theta\in\Theta$, where $\mu_m$ denotes the empirical measure defined by the subsample $X_{n-m+1},\ldots,X_n$. If the minimum does not exist, we select $\theta_n$ such that $\Delta_{\theta_n}<\inf_{\theta^*\in\Theta}\Delta_{\theta^*}+1/n$. Define $f_n=f_{n-m,\theta_n}$. Using Theorem 5.13 conditionally, we immediately obtain:

**Theorem 5.20.** *If $\int f_{n-m,\theta}=1$ for all $\theta\in\Theta$, then for the minimum distance estimate $f_n$ as defined above, we have*

$$\int|f_n-f|\le 3\inf_{\theta\in\Theta}\int|f_{n-m,\theta}-f|+4\Delta+\frac{3}{n},$$

*where*

$$\Delta=\sup_{A\in\mathcal{A}_\Theta}\left|\int_A f-\mu_m(A)\right|.$$

In order to use Theorem 5.20, first we have to make sure that $\inf_{\theta\in\Theta}\int|f_{n-m,\theta}-f|$ is not much larger than $\inf_{\theta\in\Theta}\int|f_{n,\theta}-f|$, that is, holding out $m$ samples does not hurt. In the next section we will see that for many important families of estimates, this is indeed the case. The second part of the analysis is then purely combinatorial, as upper bounds for the value of $\Delta$ may be obtained by bounding the shatter coefficients of the class $\mathcal{A}_\Theta$.

Many classical nonparametric density estimates are *additive* estimates, that is, they can be written in the form

$$g_n(x)=\frac{1}{n}\sum_{i=1}^n K(x,X_i),$$

where $K:\mathcal{R}^d\times\mathcal{R}^d\to\mathcal{R}$ is a measurable function, and $\int K(x,y)dx=1$ for all $y$. We say that the additive estimate $g_n$ is *regular* if for each $x$, $\mathbb{E}\{|K(x,X)|\}<\infty$. Examples of additive estimates include the kernel, histogram, series, and wavelet estimates. Theorem 5.21 below is a straightforward extension of a slightly less general inequality in Devroye and Lugosi (1996).

**Theorem 5.21.** *Let $\Theta$ be a class of parameters, and assume that each density estimate $f_{n,\theta}(x) = (1/n)\sum_{i=1}^{n} K_\theta(x,X_i)$ is additive and regular. Denote $J_{n,\theta} = \int |f_{n,\theta} - f|$. If $m > 0$ is a positive integer such that $2m \le n$, then*

$$\frac{\inf_{\theta\in\Theta}\mathbb{E}\{J_{n-m,\theta}\}}{\inf_{\theta\in\Theta}\mathbb{E}\{J_{n,\theta}\}} \le 1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}}.$$

This means that by decreasing the sample size to $n-m$, the performance of the best estimate in the class cannot deteriorate by more than a constant factor. If $m$ is small relative to $n$, the loss in the $L_1$ error is negligible. The proof uses some simple results about sums of independent random variables developed in Lemmas 5.4 through 5.6.

**Lemma 5.4.** *Let $X$ and $Y$ be independent random variables, and let $\mathbb{E}\{Y\} = 0$. Then $\mathbb{E}\{|X + Y|\} \ge \mathbb{E}\{|X|\}$.*

PROOF. We write $X = \mathbb{E}\{X+Y|X\}$, and use Jensen's inequality:

$$\mathbb{E}\{|X|\} = \mathbb{E}\{|\mathbb{E}\{X+Y|X\}|\} \le \mathbb{E}\{|X+Y|\}.$$

□

**Lemma 5.5.** (KHINCHINE'S INEQUALITY) *Let $a_1,\dots,a_n$ be real numbers, and let $\sigma_1,\dots,\sigma_n$ be i.i.d. sign variables with $\mathbb{P}\{\sigma_1 = 1\} = \mathbb{P}\{\sigma_1 = -1\} = 1/2$. Then*

$$\mathbb{E}\left\{\left|\sum_{i=1}^{n} a_i\sigma_i\right|\right\} \ge \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{n} a_i^2}.$$

The best constant in Khintchine's inequality is due to Szarek (1976) (see also Haagerup, 1978). The basis of the proof given here is Lemma 7.1 in Devroye and Györfi (1985).

Compare this with the closely related upper bound

$$\mathbb{E}\left\{\left|\sum_{i=1}^{n} a_i\sigma_i\right|\right\} \le \sqrt{\mathbb{E}\left\{\left(\sum_{i=1}^{n} a_i\sigma_i\right)^2\right\}} = \sqrt{\sum_{i=1}^{n} a_i^2}.$$

PROOF. The proof of the inequality may be found in Szarek (1976). Here we give a short proof with a suboptimal constant ($1/\sqrt{3}$ instead of $1/\sqrt{2}$). First note that for any random variable $X$ with finite fourth moment,

$$\mathbb{E}\{|X|\} \ge \frac{\left(\mathbb{E}\{X^2\}\right)^{3/2}}{\left(\mathbb{E}\{X^4\}\right)^{1/2}}.$$

Indeed, for any $a > 0$, the function $1/x + ax^2$ is minimal on $(0,\infty)$ when $x^3 = 1/(2a)$. Thus,

$$\frac{x + ax^4}{x^2} \geq (2a)^{1/3} + \frac{a}{(2a)^{2/3}} = {}^3\!/\!_2 (2a)^{1/3}.$$

Replace $x$ by $|X|$ and take expectations:

$$\mathbb{E}\{|X|\} \geq {}^3\!/\!_2 (2a)^{1/3}\mathbb{E}\{X^2\} - a\mathbb{E}\{X^4\}.$$

The lower bound, considered as a function of $a$, is maximized if we take $a = {}^1\!/\!_2 \left(\mathbb{E}\{X^2\}/\mathbb{E}\{X^4\}\right)^{3/2}$. Resubstitution yields the claim. Applying the above inequality for $X = \sum_{i=1}^n a_i\sigma_i$ gives

$$\mathbb{E}\left\{\left|\sum_{i=1}^n a_i\sigma_i\right|\right\} \geq \frac{\left(\sum_{i=1}^n a_i^2\right)^{3/2}}{\sqrt{\sum_{i=1}^n a_i^4 + 3\sum_{i\neq j} a_i^2 a_j^2}} \geq \frac{1}{\sqrt{3}}\sqrt{\sum_{i=1}^n a_i^2},$$

where we used $\sum_{i=1}^n a_i^4 + 3\sum_{i\neq j} a_i^2 a_j^2 \leq 3(\sum_{i=1}^n a_i^2)^2$. □

Finally, we generalize the above inequality for general sums of independent random variables:

**Lemma 5.6.** *Let $Y_1,\ldots,Y_n$ be i.i.d. zero mean random variables. Then*

$$\mathbb{E}\left\{\left|\sum_{i=1}^n Y_i\right|\right\} \geq \sqrt{\frac{n}{8}}\mathbb{E}\{|Y_1|\}.$$

PROOF. The proof uses symmetrization. We introduce the i.i.d. random variables $Y_1',\ldots,Y_n'$, distributed as the $Y_i$'s, and independent of them. Let $\sigma_1,\ldots,\sigma_n$ be i.i.d. sign variables with $\mathbb{P}\{\sigma_1 = -1\} = \mathbb{P}\{\sigma_1 = 1\} = {}^1\!/\!_2$, independent of $Y_1, Y_1',\ldots,Y_n,Y_n'$. Then

$$\begin{aligned}
\mathbb{E}\left\{\left|\sum_{i=1}^n Y_i\right|\right\} &= {}^1\!/\!_2\left(\mathbb{E}\left\{\left|\sum_{i=1}^n Y_i\right|\right\} + \mathbb{E}\left\{\left|\sum_{i=1}^n Y_i'\right|\right\}\right)\\
&\geq {}^1\!/\!_2\mathbb{E}\left\{\left|\sum_{i=1}^n (Y_i - Y_i')\right|\right\}\\
&= {}^1\!/\!_2\mathbb{E}\left\{\left|\sum_{i=1}^n \sigma_i(Y_i - Y_i')\right|\right\}\\
&= {}^1\!/\!_2\mathbb{E}\mathbb{E}\left\{\left|\sum_{i=1}^n \sigma_i(Y_i - Y_i')\right| \,\middle|\, Y_1, Y_1',\ldots,Y_n,Y_n'\right\}\\
&\geq \frac{1}{2\sqrt{2}}\mathbb{E}\sqrt{\sum_{i=1}^n (Y_i - Y_i')^2} \quad \text{(by Lemma 5.5)}\\
&\geq \frac{\sqrt{n}}{2\sqrt{2}}\mathbb{E}\{|Y_i - Y_i'|\}\\
&\geq \frac{\sqrt{n}}{2\sqrt{2}}\mathbb{E}\{|Y_i|\} \quad \text{(by Lemma 5.4).}
\end{aligned}$$

□

**Lemma 5.7.** *For any density estimate $g_n$,*

$$\mathbb{E}\int |f-g_n| \geq 1/2\,\mathbb{E}\int |g_n-\mathbb{E}g_n|.$$

PROOF. Sum the two inequalities

$$\mathbb{E}\int |f-g_n| \geq \int |f-\mathbb{E}g_n| \quad \text{(Jensen's inequality)},$$

and

$$\mathbb{E}\int |f-g_n| \geq \mathbb{E}\int |g_n-\mathbb{E}g_n| - \int |f-\mathbb{E}g_n| \quad \text{(triangle inequality)}.$$

□

PROOF OF THEOREM 5.21. Note the following:

$$\begin{aligned}
\inf_{\theta\in\Theta}\mathbb{E}\{J_{n-m,\theta}\} &\leq \inf_{\theta\in\Theta}\mathbb{E}\{J_{n,\theta}\}\times\sup_{\theta\in\Theta}\left(\frac{\mathbb{E}\{J_{n-m,\theta}\}}{\mathbb{E}\{J_{n,\theta}}\}\right)\\
&= \inf_{\theta\in\Theta}\mathbb{E}\{J_{n,\theta}\}\times\left(1+\sup_{\theta\in\Theta}\frac{\mathbb{E}\{J_{n-m,\theta}-\mathbb{E}\{J_{n,\theta}\}\}}{\mathbb{E}\{J_{n,\theta}\}}\right).
\end{aligned}$$

The supremum is rewritten as follows:

$$\begin{aligned}
\sup_{\theta\in\Theta}\frac{\mathbb{E}\{J_{n-m,\theta}-\mathbb{E}\{J_{n,\theta}\}\}}{\mathbb{E}\{J_{n,\theta}\}} &\leq \sup_{\theta\in\Theta}\frac{\mathbb{E}\{\int|f_{n-m,\theta}-f_{n,\theta}|\,dx\}}{\mathbb{E}\{J_{n,\theta}\}}\\
&\leq 2\sup_{\theta\in\Theta}\frac{\mathbb{E}\{\int|f_{n-m,\theta}-f_{n,\theta}|\,dx\}}{\mathbb{E}\{\int|f_{n,\theta}-\mathbb{E}f_{n,\theta}|\,dx\}},
\end{aligned}$$

where we used Lemma 5.7. Fix $x$ and $\theta$ for now. Introduce

$$Y_i = K_\theta(x,X_i)-\mathbb{E}\{K_\theta(x,X)\},$$

and denote the partial sums of $Y_i$'s by $S_j = Y_1+\cdots+Y_j$. By assumption, for fixed $x$ and $\theta$, the first absolute moment of $Y_1$ is finite. Then observe the following:

$$n|f_{n-m,\theta}-f_{n,\theta}| = \left|\frac{m}{n-m}(Y_1+\cdots+Y_{n-m})-(Y_{n-m+1}+\cdots+Y_n)\right|$$

so that

$$\mathbb{E}\left\{n|f_{n-m,\theta}-f_{n,\theta}|\right\} \leq \frac{m}{n-m}\mathbb{E}\{|S_{n-m}|\}+\mathbb{E}\{|S_m|\}.$$

Also, $n|f_{n,\theta} - \mathbb{E}f_{n,\theta}| = |S_n|$, which implies $\mathbb{E}\left\{n|f_{n,\theta} - \mathbb{E}f_{n,\theta}|\right\} = \mathbb{E}\{|S_n|\}$. Still holding $x$ and $\theta$ fixed, we bound the following ratio:

$$\begin{aligned}
\frac{\mathbb{E}\{|f_{n-m,\theta} - f_{n,\theta}|\}}{\mathbb{E}\{|f_{n,\theta} - \mathbb{E}f_{n,\theta}|\}} &\leq \frac{(m/(n-m)\mathbb{E}\{|S_{n-m}|\} + \mathbb{E}\{|S_m|\}}{\mathbb{E}\{|S_n|\}} \\
&\leq \frac{m}{n-m} + \frac{\mathbb{E}\{|S_m|\}}{\mathbb{E}\{|S_n|\}} \\
&\quad \text{(because } \mathbb{E}\{|S_n|\} \geq \mathbb{E}\{|S_{n-m}|\}\text{)} \\
&\leq \frac{m}{n-m} + \frac{\mathbb{E}\{|S_m|\}}{\sqrt{\lfloor n/m \rfloor/8}\,\mathbb{E}\{|S_m|\}} \quad \text{(by Lemmas 5.4 and 5.6)} \\
&\leq \frac{m}{n-m} + 4\sqrt{\frac{m}{n}} \quad \text{(if } 2m \leq n\text{).}
\end{aligned}$$

This implies that for any fixed $\theta$:

$$\mathbb{E}\int |f_{n-m,\theta} - f_{n,\theta}|\,dx \leq \left(\frac{m}{n-m} + 4\sqrt{\frac{m}{n}}\right) \mathbb{E}\int |f_{n,\theta} - \mathbb{E}f_{n,\theta}|\,dx.$$

The result now follows without work. □

Combining Theorems 5.20, 5.21 and 1.9, we readily obtain:

**Theorem 5.22.** *Let the set* $\Theta$ *determine a class of regular additive density estimates with* $\int f_{n-m,\theta} = 1$ *for all* $\theta \in \Theta$*. Then for all* $n$*,* $m \leq n/2$*,* $\Theta$*, and* $f$*:*

$$\begin{aligned}
\mathbb{E}\int |f_n - f| \quad &\leq \quad 3 \inf_{\theta\in\Theta} \mathbb{E}\int |f_{n,\theta} - f| \left(1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}}\right) \\
&\quad + 8\mathbb{E}\left\{\sqrt{\frac{\log 2\mathbb{S}_{\mathcal{A}_\Theta}(m)}{m}}\right\} + \frac{3}{n}.
\end{aligned}$$

We note that $\mathbb{S}_{\mathcal{A}_\Theta}(m)$ is a random variable that depends upon $X_1, \ldots, X_{n-m}$, since the definition of $\mathcal{A}_\Theta$ involves these data points. In most applications, $\mathcal{A}_\Theta$ can be bounded uniformly over all values of $X_1, \ldots, X_{n-m}$.

EXAMPLE: HISTOGRAM ESTIMATE. The simplest additive estimates are histograms. A histogram density estimate based on a partition $P$ of $\mathcal{R}^d$ is defined by

$$f_{n,P}(x) = \frac{\mu_n(A_P(x))}{\lambda(A_P(x))},$$

where $\mu_n$ is the empirical measure based on the sample $X_1, \ldots, X_n$, $\lambda$ is the Lebesgue measure on $\mathcal{R}^d$, $A_P(x)$ denotes the cell of the partition $P$ into which $x$ falls, and $a/\infty$ is defined to be zero.

As a first application of Theorem 5.22, we consider the selection of a partition for a histogram. Formally, let $\mathcal{P}$ be a family of partitions of $\mathcal{R}^d$, and to each partition $P \in \mathcal{P}$, assign the corresponding histogram estimate $f_{n,P}(x)$. We use the minimum distance estimate based on data splitting to select a partition from the class $\mathcal{P}$, thus obtaining the density estimate $f_n$. First observe that any histogram estimate is regular and additive. However, if $P$ contains a cell with infinite Lebesgue measure, then $f_{n,P}$ does not necessarily integrate to one, which makes the use of Theorem 5.22 illegitimate. To fix this problem, we map $\mathcal{R}^d$ to $[0,1]^d$, and consider only partitions of the cube.

To apply Theorem 5.22, we merely need to obtain upper bounds for the shatter coefficients $\mathbb{S}_{\mathcal{A}_{\mathcal{P}}}(m)$, where $\mathcal{A}_{\mathcal{P}}$ is the Yatracos class of all sets of the form

$$\{x : f_{n-m,P_1}(x) > f_{n-m,P_2}(x)\}, \qquad P_1, P_2 \in \mathcal{P}.$$

As a prototype example, consider the class $\mathcal{P}$ of partitions of $\mathcal{R}$ into intervals of length $h$, where $h = 2^k$ for some integer $k$. The estimate $f_{n,P}$ based on such a partition is called a regular histogram estimate. To be specific, assume that each partition is anchored at the origin, that is, 0 lies on the boundary of two cells. The parameter $k$ may take infinitely many possible values, and our goal is to select a close-to-optimal value. The following combinatorial argument guarantees, via Theorem 5.22, that the minimum distance estimate may be successfully used for the selection of $h$ from a dyadic collection of interval lengths.

**Lemma 5.8.**

$$\mathbb{S}_{\mathcal{A}_{\mathcal{P}}}(m) \le (m+1)n^2.$$

*Also, for all $n$, $m \le n/2$, and $f$, if $f_n$ is the regular histogram estimate picked by the minimum distance method for $\mathcal{P}$, we have*

$$\begin{aligned} \mathbb{E}\int |f_n - f| \quad &\le \quad 3 \inf_{\theta \in \mathcal{P}} \mathbb{E}\int |f_{n,\theta} - f| \left(1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}}\right) \\ &\qquad + 8\sqrt{\frac{\log(2(m+1)n^2)}{m}} + \frac{3}{n}. \end{aligned}$$

PROOF. Each partition $P_k$ in $\mathcal{P}$ is indexed by a (possibly negative-valued) integer $k$. Let $A_k(x)$ be the cell of $P_k$ containing $x$. Observe that the value of the vector

$$z_k = \left(\sum_{i=1}^{n-m} \mathbb{I}_{[y_1 \in A_k(X_i)]}, \ldots, \sum_{i=1}^{n-m} \mathbb{I}_{[y_m \in A_k(X_i)]}\right)$$

can take at most $n$ different values when we vary $k$ but keep the $y_i$'s and $X_i$'s fixed. To see this, observe that there exists a large integer $k_0$ such that all data points fall in the same cell of $P_{k_0}$. Then for all $k \ge k_0$ the value of $z_k$ is the same. As $k$ decreases in steps of size one, $z_k$ can only change when there exists a cell $[a,b)$ of $P_k$ such that $[a,(a+b)/2)$ and $[(a+b)/2,b)$ both contain at least one data point among $\{X_1,\ldots,X_{n-m},y_1,\ldots,y_m\}$. But this can happen for at most $n-1$ different values of $k$. Thus,

$$|\{(z_k,z_\ell) : k,\ell \in \{0,\pm 1,\pm 2,\ldots\}\}| \le n^2.$$

If $w$ and $w'$ are two possible values of the vector $z_k$, let $U_{w,w'}$ be the set of all pairs $(k,\ell)$ such that $(z_k,z_\ell) = (w,w')$. Fix $w = (w_1,\ldots,w_m)$ and $w' = (w'_1,\ldots,w'_m)$. If $(k,\ell) \in U_{w,w'}$, then $y_i \in A_{k,\ell} = \{x : f_{n-m,k}(x) > f_{n-m,\ell}(x)\}$ if and only if

$$\frac{w_i}{2^k} > \frac{w'_i}{2^\ell}.$$

Therefore,

$$\begin{aligned}
&|\{\{y_1,\ldots,y_m\} \cap A_{k,\ell} : (k,\ell) \in U_{(w,w')}\}| \\
&\qquad \le |\{(\mathbb{I}_{[w_1 \ge c w'_1]},\ldots,\mathbb{I}_{[w_m \ge c w'_m]}) : c > 0\}| \le m+1.
\end{aligned}$$

Thus, $\mathbb{S}_{\mathcal{A}_\mathcal{P}}(m) \le (m+1)n^2$. □

For various results on the regular histogram estimate we refer to Devroye and Györfi (1985). Consistency of the histogram estimate based on data-dependent partitions is investigated in Lugosi and Nobel (1996). Selection of the partition based on penalized maximum likelihood methods is studied by Barron, Birgé, and Massart (1999) and Castellan (2000).

## 5.8 Bandwidth Selection for Kernel Estimates

This section is about the choice of the bandwidth (or smoothing factor) $h \in (0,\infty)$ of the standard kernel estimate

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

We assume that $K : \mathcal{R}^d \to \mathcal{R}$ is a fixed kernel with $\int K = 1$. We study the minimum distance estimate based on data-splitting as described in the previous chapter, that is, we fix $m \le n/2$, define a class of densities $\mathcal{F} = \{f_{n-m,h} : h > 0\}$, and project the empirical measure $\mu_m$ (defined on the held-out data $X_{m+1},\ldots,X_n$) on $\mathcal{F}$ by the minimum distance method. Clearly, every estimate in $\mathcal{F}$ is additive, so if, in addition, $K$ is a bounded function, then Theorem 5.22 applies. To apply the bound of this theorem, we need to compute useful upper bounds for the shatter coefficient $\mathbb{S}_\mathcal{A}(m)$ of the class

$$\mathcal{A} \stackrel{\text{def}}{=} \{\{x : f_{n-m,h}(x) > f_{n-m,h'}(x)\} : \ h,h' > 0\}.$$

Of course, this depends on $n$, $m$, and the kernel function $K$. Obtaining meaningful (i.e., polynomial in $m$ and $n$) upper bounds may be difficult, or even impossible for general kernels. Thus, we begin our study with estimates based on simple kernels. The material of this section is based on Devroye and Lugosi (1997). We will not discuss implementation issues, for which we refer to Devroye (1997).

We consider kernels of the form

$$K(x) = \sum_{i=1}^{k} \alpha_i \mathbb{I}_{A_i}(x),$$

where $k < \infty$, $\alpha_1, \ldots, \alpha_k \in \mathcal{R}$, and $A_1, \ldots, A_k$ are Borel sets in $\mathcal{R}^d$ with the following property: the intersection of an infinite ray $\{x : x = tx_0, t \geq 0\}$, anchored at the origin, with any $A_i$ is an interval.

Observe that all convex sets and all star-shaped sets satisfy the above requirement. (A set $A$ is called star-shaped if $x \in A$ implies $\lambda x \in A$ for all $\lambda \in [0,1]$.) The $A_i$'s need not be disjoint. Kernels of this type are called *Riemann kernels* of parameter $k$. Important special cases include the uniform densities on ellipsoids, balls, and hypercubes. The key combinatorial result is the following:

**Lemma 5.9.** *Let $K = \sum_{i=1}^{k} \alpha_i \mathbb{I}_{A_i}$ be a Riemann kernel of parameter $k$. For all $m \geq 1$ we have*

$$\mathbb{S}_{\mathcal{A}}(m) \leq (m+1)(1+2km(n-m))^2.$$

PROOF. Let $r = n - m$. Define the vector

$$z_u = \left( \sum_{i=1}^{r} K\left(\frac{y_1 - X_i}{u}\right), \ldots, \sum_{i=1}^{r} K\left(\frac{y_m - X_i}{u}\right) \right) \in \mathcal{R}^m.$$

As we increase $u$ from zero to infinity, each component of $z_u$ changes every time $(y_j - X_i)/u$ enters or leaves a set $A_q$, $1 \leq q \leq k$ for some $X_i$, $1 \leq i \leq r$. Note that for fixed $(y_j - X_i)$, the evolution is along an infinite ray anchored at the origin. By our assumption on the possible form of the sets $A_q$, the number of different values a component can take in its history (as $u \uparrow \infty$) is bounded by $2kr$. As there are $m$ components, the cardinality of the set of different values of $z_u$ is bounded as

$$|\{z_u : u > 0\}| \leq 1 + 2kmr.$$

Thus,

$$|\{(z_u, z_v) : u, v > 0\}| \leq (1 + 2kmr)^2.$$

Let $\mathcal{W} = \{(w, w') : (w, w') = (z_u, z_v) \text{ for some } u, v > 0\}$. For fixed $(w, w') \in \mathcal{W}$, let $U_{(w,w')}$ denote the collection of all $(u, v)$ such that $(z_u, z_v) = (w, w')$. For $(u, v) \in U_{(w,w')}$, we have

$$y_i \in A_{u,v} \quad \text{if and only if} \quad w_i \geq \left(\frac{u}{v}\right)^d w_i',$$

where $w, w'$ have components $w_i, w'_i$, respectively, $1 \le i \le m$. Thus,

$$\begin{aligned}
&|\{\{y_1,\dots,y_m\} \cap A_{u,v} : (u,v) \in U_{(w,w')}\}| \\
&\qquad \le \left|\left\{\left(\mathbb{I}_{[w_1 \ge cw'_1]},\dots,\mathbb{I}_{[w_m \ge cw'_m]}\right) : c \ge 0\right\}\right| \le m+1.
\end{aligned}$$

But then

$$\begin{aligned}
&|\{\{y_1,\dots,y_m\} \cap A_{u,v} : (u,v) > 0\}| \\
&\qquad \le (m+1)|U_{(w,w')}| \le (m+1)(1+2kmr)^2.
\end{aligned}$$

□

For Riemann kernels, Lemma 5.9 is the last missing link. Plugging the upper bound of the lemma into Theorem 5.22, we obtain a nonasymptotic, density-free inequality.

Most kernel functions used in practice are not Riemann kernels, and bounding the complexity of the Yatracos class based on such kernels is difficult. Luckily, most kernels can be well approximated by Riemann kernels, and this suggests the following solution for the bandwidth selection problem: first select a positive integer $k$ and a Riemann kernel $K' = \sum_{i=1}^{k} \alpha_i \mathbb{I}_{A_i}$ such that

$$\int |K - K'| \le \frac{1}{n}.$$

Note that this is always possible if $K$ is Riemann integrable. Now define the kernel estimates using the approximating kernel

$$f'_{n-m,h}(x) = \frac{1}{n-m} \sum_{i=1}^{n-m} K'_h(x - X_i)$$

for all $h > 0$. Finally, select the smoothing factor $H$ by the minimum distance estimate over this class, and define the density estimate

$$f_n(x) = \frac{1}{n-m} \sum_{i=1}^{n-m} K_H(x - X_i).$$

This $H$ is called the Riemann approximation bandwidth and $f_n$ is the Riemannian kernel estimate. The size of the smallest $k$ for which this is possible depends on the kernel $K$. We call this the *kernel complexity* $\kappa_n$ of $K$:

$$\kappa_n = \min\left\{k : \exists K' = \sum_{i=1}^{k} \alpha_i \mathbb{I}_{A_i} \text{ such that } \int |K - K'| \le \frac{1}{n}\right\}.$$

The main result of this section is the following performance bound for $f_n$:

**Theorem 5.23.** *Let $K$ be a bounded kernel with kernel complexity $\kappa_n$, and let $m \le n/2$. Then for all densities $f$, the Riemannian kernel estimate $f_n$ satisfies*

$$\begin{aligned}\mathbb{E}\int |f_n - f| \quad &\le \quad 3\left(1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}}\right) \inf_h \mathbb{E}\int |f_{n,h} - f| \\ &\qquad + 8\sqrt{\frac{\log\left(2(m+1)(1+2\kappa_n m(n-m))^2\right)}{m}} + \frac{31}{n}.\end{aligned}$$

PROOF. Observe that for each $h$, $\int |f_{n-m,h} - f'_{n-m,h}| \le \int |K - K'|$. Thus,

$$\begin{aligned}&\mathbb{E}\int |f_n - f| \\ &\le \mathbb{E}\int |f'_{n-m,H} - f| + \int |K - K'| \\ &\le 3\left(1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}}\right) \inf_h \mathbb{E}\int |f'_{n,h} - f| \\ &\qquad + 8\sqrt{\frac{\log\left(2(m+1)(1+2\kappa_n m(n-m))^2\right)}{m}} + \int |K - K'| + \frac{3}{n} \\ &\qquad \text{(by the results of the previous section)} \\ &\le 3\left(1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}}\right) \inf_h \mathbb{E}\int |f_{n,h} - f| \\ &\qquad + 8\sqrt{\frac{\log\left(2(m+1)(1+2\kappa_n m(n-m))^2\right)}{m}} \\ &\qquad + \left(1 + 3\left(1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}}\right)\right) \int |K - K'| + \frac{3}{n}.\end{aligned}$$

□

To understand the implications of Theorem 5.23, consider the simplest, though suboptimal choice $m = \lfloor n/2 \rfloor$. Then we obtain

$$\mathbb{E}\int |f_n - f| \le 43 \inf_h \mathbb{E}\int |f_{nh} - f| + c\sqrt{\frac{\log(n\kappa_n)}{n}} + \frac{31}{n},$$

where $c$ is a universal constant, independent of $f$ and $K$. Taking $m$ as a smaller fraction of $n$, we may decrease the factor 43 to close to 3, at the expense of increasing the factor $c$ in front of the second term. As we will see later, for most kernels the first term converges to zero much slower than $\sqrt{\log n/n}$ for *all* densities. Therefore, the first term asymptotically dominates the second one if the kernel complexity $\kappa_n$ is a polynomial function of $n$. In the next section we show that this is the case for all important kernels.

KERNEL COMPLEXITY: UNIVARIATE EXAMPLES. In this section we provide a list of examples of kernels on $\mathcal{R}$ whose complexity $\kappa_n$ is bounded by a polynomial of $n$. Let us warm up with two toy examples:

UNIFORM KERNELS. If $K(x) = \mathbb{I}_A(x)$ for a star-shaped set $A$, then obviously $\kappa_n = 1$ for all $n > 1$.

ISOSCELES TRIANGULAR DENSITY. If $K(x) = (1 - |x|)_+$, then elementary calculation shows that for all $n$, $\kappa_n \le n + 1$.

Most important kernels are nonnegative, unimodal, and symmetric about the origin. For such kernels, the following bound is useful.

SYMMETRIC UNIMODAL KERNELS. Consider a symmetric unimodal density (i.e., $K \ge 0$ and $\int K = 1$) on the real line. Let $\beta$ be the last positive value for which $\int_\beta^\infty K \le 1/(4n)$. Partition $[0,\beta]$ and $[-\beta,0]$ into $N = \lceil 4nK(0)\beta \rceil$ equal intervals. On each interval, let $K'$ be constant with value equal to the average of $K$ over that interval. Let $\gamma = \int_\beta^\infty K/K(\beta)$, and set $K'(x) = K(\beta)$ on $[\beta, \beta+\gamma]$ and $[-\beta-\gamma, -\beta]$. Note that $\int K' = 1$, $\int |K - K'| \le 1/n$, and that $K'$ is Riemann with parameter $k \le 2N + 2 \le 8nK(0)\beta + 10$. Thus, $\kappa_n \le 8nK(0)\beta + 10$.

EXAMPLE 1 (BOUNDED COMPACT SUPPORT KERNELS). If $K(x) \le a\mathbb{I}_{[-b,b]}(x)$ and $K$ is symmetric, nonnegative, and unimodal (such as the Epanechnikov kernel), then $\kappa_n \le 8nab + 10$.

EXAMPLE 2 (THE NORMAL KERNEL). When $K(x) = e^{-x^2/2}/\sqrt{2\pi}$, we have $K(0) = 1/\sqrt{2\pi}$. Since for $\beta \ge 1$,

$$\int_\beta^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx \le \frac{1}{\sqrt{2\pi}}\frac{1}{\beta} e^{-\beta^2/2} \le \frac{1}{\sqrt{2\pi}} e^{-\beta^2/2},$$

we may take $\beta = \sqrt{2\log(4n/\sqrt{2\pi})}$. Thus, for all $n > 1$:

$$\kappa_n \le \frac{8n\sqrt{\log n}}{\sqrt{\pi}} + 10.$$

EXAMPLE 3 (THE CAUCHY KERNEL). Take $K(x) = 1/(\pi(1+x^2))$. Note that $K(0) = 1/\pi$, and that $\beta = \pi/(4n)$ will do. Therefore,

$$\kappa_n \le \frac{32n^2}{\pi^2} + 10.$$

EXAMPLE 4 (KERNELS WITH POLYNOMIAL TAILS). Note that if $K$ is a symmetric unimodal density, and $|K(x)| \le c/(1+|x|^{\gamma+1})$ for some $c < \infty$, $\gamma > 0$, then $\kappa_n = O(n^{1+1/\gamma})$. In fact, for most cases of interest, $\kappa_n = O(n^\alpha)$ for some finite constant $\alpha > 0$.

Sometimes it may be beneficial to use kernels which may take negative values. The next bound will be useful even for kernels that oscillate infinitely many times.

LIPSCHITZ KERNELS WITH A POSSIBLY HEAVY TAIL. Let $K$ be a univariate kernel that is Lipschitz with Lipschitz constant $C$, and assume that $|K(x)| \leq D/x^2$ for another constant $D$. Then

$$\kappa_n \leq 1 + 32CD^2n^3.$$

PROOF. Take $r = 4Dn$ and note that $\int_{|x|>r} |K| \leq 2D/r = 1/(2n)$. Partition $[-r, r]$ into $q$ equal intervals of length $2r/q$ each. Define a Riemann kernel $K'$ of order $q$ taking a constant value on each of these intervals, equal to the average of $K$ over the intervals. By the Lipschitz condition, on any such interval $A$, $\int_A |K - K'| \leq C(2r/q)^2/2 = 2Cr^2/q^2$. Let $\gamma^+ = (1/K(r)) \int_r^\infty K$ and $\gamma^- = (1/K(-r)) \int_{-\infty}^{-r} K$, and set $K'(x) = K(r)$ on $[r, r+\gamma^+]$ and $K'(x) = K(-r)$ on $[-r-\gamma^-, r]$. Thus, $\int K' = 1$ and

$$\int |K - K'| \leq \frac{1}{2n} + \frac{2Cr^2}{q} \leq \frac{1}{n}$$

provided that $q \geq 2Cr^2n = 32CD^2n^3$. □

Finally, we mention a huge class of kernels, containing nearly every one-dimensional kernel.

KERNELS OF BOUNDED VARIATION. If $K$ is symmetric and a difference of two monotone functions, that is, $K = K_1 - K_2$, $K_1 \downarrow 0$, $K_2 \downarrow 0$ on $[0, \infty)$, then each $K_1, K_2$ may be approximated as above. Thus, in particular, if $K$ is of *bounded variation*, and $|K(x)| \leq c/(1 + |x|^{\gamma+1})$ for some $c < \infty$, $\gamma > 0$, then we may approximate with $\kappa_n = O(n^{1+1/\gamma})$.

PRODUCT KERNELS. If $K = K_1 \times \cdots \times K_d$ is a product of $d$ univariate kernels, and if we approximate $K_i$ with $K_i'$ with parameter $\kappa_{nd}^{(i)}$ for all $i$ (where $\kappa_{nd}^{(i)}$ is the kernel complexity of $K_i$ of precision $1/(nd)$), and form $K' = K_1' \times \cdots \times K_d'$, then $K'$ is a weighted sum of indicators of product sets, and it is Riemann with parameter not exceeding $\prod_{i=1}^d \kappa_{nd}^{(i)}$. Furthermore,

$$\begin{aligned}
\int |K - K'| &\leq \int |K_1 \times \cdots \times K_{d-1} \times K_d - K_1 \times \cdots \times K_{d-1} \times K_d'| \\
&\quad + \ldots \\
&\quad + \int |K_1 \times K_2' \cdots \times K_d' - K_1' \times K_2' \cdots \times K_d'| \\
&\leq d\left(\frac{1}{nd}\right) \\
&= \frac{1}{n}.
\end{aligned}$$

Thus, $\kappa_n$ is bounded by $\prod_{i=1}^d \kappa_{nd}^{(i)}$.

KERNELS THAT ARE FUNCTIONS OF $\|x\|$. Assume that $K(x) = M(\|x\|)$, where $M$ is a bounded nonnegative monotone decreasing function on $[0,\infty)$. Then we may approximate $M$ by a stepwise constant function $M'$, and use the Riemann kernel $K'(x) = M'(\|x\|)$ in the estimate as an approximation of $K$. Clearly,

$$\int |K(x) - K'(x)|\,dx = \int_0^\infty c_d u^{d-1} |M(u) - M'(u)|\,du,$$

where $c_d$ is $d$ times the volume of the unit ball in $\mathcal{R}^d$. We may define $M'$ as follows. Let $\beta$ be the largest positive number for which $\int_\beta^\infty c_d u^{d-1} M(u)\,du \le 1/(2n)$. Partition $[0,\beta]$ into $N = \lceil 2nc_d M(0)\beta^d \rceil$ equal intervals. On each interval, let $M'$ equal to the average of $M$ over that interval. Let $\gamma = \int_\beta^\infty c_d u^{d-1} M(u)\,du / M(\beta)$, and set $M'(u) = M(\beta)$ on $u \in [\beta, \beta+\gamma]$, and let $M'(u) = 0$ for $u > \gamma$. Clearly $\int K' = 1$, and that $K'$ is Riemann with parameter $k = N + 1 \le 2nc_d K(0)\beta^d + 2$. Moreover,

$$\begin{aligned}
&\int |K(x) - K'(x)|\,dx \\
&\quad = \int_0^\beta c_d u^{d-1} |M(u) - M'(u)|\,du + \int_\beta^\infty c_d u^{d-1} |M(u) - M'(u)|\,du \\
&\quad \le \frac{1}{2n} + c_d \beta^{d-1} \int_0^\beta |M(u) - M'(u)|\,du \\
&\quad \le \frac{1}{2n} + c_d \beta^{d-1} \frac{M(0)\beta}{N} \\
&\quad \le \frac{1}{n}.
\end{aligned}$$

Thus,

$$\kappa_n \le 2nc_d M(0)\beta^d + 2.$$

THE MULTIVARIATE STANDARD NORMAL KERNEL. We may apply the bound of the previous paragraph to the multivariate normal density. First note that it suffices to take $\beta = 2\sqrt{2\log n}$. From this, we deduce that the kernel complexity is

$$\kappa_n = O(n \log^{d/2} n).$$

ASYMPTOTIC OPTIMALITY. One important corollary of Theorem 5.23 is that asymptotically the error of the estimate stays within a factor of three of that of the kernel estimate with the best possible smoothing factor.

**Theorem 5.24.** (DEVROYE AND LUGOSI, 1996, 1997) *Let $K$ be a bounded nonnegative kernel on the real line with complexity $\kappa_n$ bounded by some polynomial of $n$. If $m/n \to 0$ and*

*$m/(n^{4/5}\log n) \to \infty$ as $n \to \infty$, then for all densities $f$ on $\mathcal{R}$:*

$$\sup_f \limsup_{n\to\infty} \frac{\mathbb{E}\{\int |f_n - f|\}}{\inf_h \mathbb{E}\{\int |f_{n,h} - f|\}} \le 3.$$

The statement is an easy consequence of the fact that by Theorem 5.23, with the given choice of $m$,

$$\mathbb{E}\int |f_n - f| \le (3 + o(1)) \inf_h \mathbb{E}\int |f_{nh} - f| + o(n^{-2/5}),$$

and the following lower bound due to Devroye and Penrod (1984).

**Lemma 5.10.** *Let $K$ be a nonnegative kernel on the real line. Then for any density $f$:*

$$\liminf_{n\to\infty} n^{2/5} \inf_h \mathbb{E}\int |f_{nh} - f| \ge 0.86.$$

## 5.9 References


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

Akaike, H. (1954). "An approximation to the density function," *Annals of the Institute of Statistical Mathematics*, vol. 6, pp. 127–132.

Anthony, M. and Bartlett, P. L. (1999). *Neural Network Learning: Theoretical Foundations*, Cambridge University Press, Cambridge.

Azuma, K. (1967). "Weighted sums of certain dependent random variables," *Tohoku Mathematical Journal*, vol. 37, pp. 357–367.

Barron, A. R. (1988). The convergence in information of probability density estimates. *IEEE ISIT, Kobe, Japan*.

Barron, A. R. (1989). Uniformly powerful goodness of fit tests. *Ann. Statist.* 17, 107-124.

Barron, A. Birgé, L. and Massart, P. (1999). "Risk bounds for model selection via penalization," *Probability Theory and Related Fields*, vol. 113, pp. 301–415.

Barron, A. R., Györfi, L. and van der Meulen, E. C. (1992). Distribution estimates consistent in total variation and in two types of information divergence. *IEEE Trans. on Information Theory*, 38, pp. 1437-1454.

Beirlant, J., Berlinet, A. and Györfi, L. (1999). On piecewise linear density estimation. *Statistica Neerlandica*, 53, pp. 287-308.

Beirlant, J., Devroye, L., Györfi, L. and Vajda I. (2001). Large deviations of divergence measures on partitions. *J. Statistical Planning and Inference*, 93, pp. 1-16.

Beirlant, J. and Györfi, L. (1998). On the $L_1$-error in histogram density estimation: the multidimensional case. *J. Nonparametric Statistics*, 9, pp. 197-216.

Beirlant, J., Györfi, L. and Lugosi, G. (1994). On the asymptotic normality of the $L_1$- and $L_2$- errors in the histogram density estimation. *Canadian J. Statistics*, 22, pp. 309-318.

Beirlant, J. and Mason, D. M. (1995). "On the asymptotic normality of $L_p$-norms of empirical functionals," *Mathematical Methods of Statistics*, vol. 4, pp. 1–19.

Berlinet, A. (1995). Central limit theorems in functional estimation. *Bulletin of the International Statistical Institute*, 56, pp. 531-548.

Berlinet, A. and Devroye, L. (1994). "A comparison of kernel density estimates," *Publications de l'Institut de Statistique de l'Université de Paris*, vol. 38, pp. 3–59.

Berlinet, A., Devroye, L. and Györfi, L. (1995). Asymptotic normality of $L_1$ error in density estimation. *Statistics*, 26, pp. 329-343.

Berlinet, A., Györfi, L. and van der Meulen, E. (1995). The asymptotic normality of relative entropy in multivariate density estimation. *Publications de l'Institut de Statistique de l'Université de Paris*, 41, pp. 3-27.

Birgé, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probability Theory and Related Fields*, 71, pp. 271-291.

Bretagnolle, J. and Huber, C. (1979). "Estimation des densités: Risque minimax," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 47, pp. 119–137.

Broniatowski, M., Deheuvels, P. and Devroye, L. (1989). "On the relationship between stability of extreme order statistics and convergence of the maximum likelihood kernel density estimate," *Annals of Statistics*, vol. 17, pp. 1070–1086.

Butzer, P. L. and Nessel, R. J. (1971). *Fourier Analysis and Approximation*, Birkhäuser-Verlag, Basel, 1971.

Cao, R., Cuevas, A. and González–Manteiga, W. (1994). "A comparative study of several smoothing methods in density estimation," *Computational Statistics and Data Analysis*, vol. 17, pp. 153–176.

Castellan, G. (2000). “Sélection d’histogrammes ou de modèles exponentiels de polynomes par morceaux à l’aide d’un critère de type Akaike,” Thèse, Mathématiques, Université de Paris-Sud.

Cline, D. B. H. (1988). “Admissible kernel estimators of a multivariate density,” *Annals of Statistics*, vol. 16, pp. 1421–1427.

Cline, D. B. H. (1990) “Optimal kernel estimation of densities,” *Annals of the Institute of Statistical Mathematics*, vol. 42, pp. 287–303.

Csiszár, I. (1967). Information-type measures of divergence of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2, pp. 299-318.

Csiszár, I. and Körner, J. (1981). *Information Theory: Coding Theorems for Memoryless Systems*. Academic Press, New York.

Csörgő, M. and Horváth, L. (1988). “Central limit theorems for $L_p$-norms of density estimators,” *Probability Theory and Related Fields*, vol. 80, pp. 269–291.

Deheuvels, P. (1977). “Estimation nonparamétrique de la densité par histogrammes generalisés,” *Publications de l’Institut de Statistique de l’Université de Paris*, vol. 22, pp. 1–23.

Dembo, A. and Zeitouni, O. (1992). *Large Deviations Techniques and Applications*. Jones and Bartlett Publishers.

Devroye, L. (1983a). The equivalence of weak, strong and complete convergence in L1 for kernel density estimates. *Annals of Statistics*, 11, pp. 896–904.

Devroye, L. (1983b). On arbitrary slow rates of global convergence in density estimation. *Zeitschrift für Wahscheinlichkeitstheorie und verwandte Gebiete*, 62, pp. 475-483.

Devroye, L. (1987). *A Course in Density Estimation*, Birkhäuser-Verlag, Boston.

Devroye, L. (1988a). The kernel estimate is relatively stable. *Probability Theory and Related Fields*, 77, pp. 521–536.

Devroye, L. (1988b) “Asymptotic performance bounds for the kernel estimate,” *Annals of Statistics*, vol. 16, pp. 1162–1179.

Devroye, L. (1989a) “The double kernel method in density estimation,” *Annales de l’Institut Henri Poincaré*, vol. 25, pp. 533–580.

Devroye, L. (1989b). “Nonparametric density estimates with improved performance on given sets of densities,” *Statistics (Mathematische Operationsforschung und Statistik)*, vol. 20, pp. 357–376.

Devroye, L. (1991). “Exponential inequalities in nonparametric estimation,” in: *Nonparametric Functional Estimation and Related Topics*, (edited by G. Roussas), pp. 31–44, NATO ASI Series, Kluwer Academic, Dordrecht.

Devroye, L. (1997). “Universal smoothing factor selection in density estimation: Theory and practice (with discussion),” *Test*, vol. 6, pp. 223–320.

Devroye, L. (1995). Another proof of a slow convergence result of Birgé. *Statistics and Probability Letters*, 23, pp. 63-67.

Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: the $L_1$ View*. Wiley.

Devroye, L. and Györfi, L. (1990). No empirical measure can converge in total variation sense for all distributions. *Annals of Statistics*, 18, pp. 1496-1499.

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer Verlag.

Devroye, L. and Lugosi, G. (1996). “A universally acceptable smoothing factor for kernel density estimation,” *Annals of Statistics*, vol. 24, pp. 2499–2512.

Devroye, L. and Lugosi, G. (1997). “Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes,” *Annals of Statistics*, vol. 25, pp. 2626–2637.

Devroye. L. and Lugosi, G. (2001). *Combinatorial Methods in Density Estimation*, Springer-Verlag, New York.

Devroye, L. and Penrod, C. S. (1984). “Distribution-free lower bounds in density estimation,” *Annals of Statistics*, vol. 12, pp. 1250–1262.

Devroye L. and Wand, M. P. (1993). “On the effect of density shape on the performance of its kernel estimate,” *Statistics*, vol. 24, pp. 215–233.

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

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

Dvoretzky, A. Kiefer, J. and Wolfowitz, J. (1956). "Asymptotic minimax character of a sample distribution function and of the classical multinomial estimator," *Annals of Mathematical Statistics*, vol. 33, pp. 642–669.

Epanechnikov, V. A. (1969). "Nonparametric estimation of a multivariate probability density," *Theory of Probability and its Applications*, vol. 14, pp. 153–158.

Goldberg P. and Jerrum, M. (1995). "Bounding the Vapnik-Chervonenkis dimension of concept classes parametrized by real numbers," *Machine Learning*, vol. 18, pp. 131–148.

Györfi, L., Páli, I. and van der Meulen, E. C. (1994). There is no universal source code for infinite alphabet. *IEEE Trans. on Information Theory*, 40, pp. 267-271.

Györfi, L. and van der Meulen, E. C. (1994). There is no density estimate consistent in information divergence for all densities. *Transactions of the Twelfth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, pp. 88-90.

Haagerup, U. (1978). "Les meilleures constantes de l'inégalité de Khintchine," *Comptes Rendus des Séances de l'Académie des Sciences de Paris. Séries A*, vol. 286, pp. 259–262.

Hall, P. and Wand, M. P. (1988). "Minimizing $L_1$ distance in nonparametric density estimation," *Journal of Multivariate Analysis*, vol. 26, pp. 59–88.

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

Holmström, L. and Klemelä, J. (1992). "Asymptotic bounds for the expected $L_1$ error of a multivariate kernel density estimator," *Journal of Multivariate Analysis*, vol. 40, pp. 245–255.

Kemperman, J. H. B. (1969). An optimum rate of transmitting information. *Ann. Math. Statist.*, 40, pp. 2156-2177.

Karpinski, M. and Macintyre, A. (1997). "Polynomial bounds for VC dimension of sigmoidal and general pfaffian neural networks," *Journal of Computer and System Science*, vol. 54, pp. 169–176.

Khovanskii, A. G. (1991). "Fewnomials," in: *Translations of Mathematical Monographs*, vol. 88, American Mathematical Society, Providence, RI.

Koiran, P. and Sontag, E. D. (1997). "Neural networks with quadratic VC dimension," *Journal of Computer and System Science*, vol. 54, pp. 190–198.

Krzyżak, A. and Linder, T. (1998). "Radial basis function networks and complexity regularization in function learning," *IEEE Transactions on Neural Networks*, vol. 9, pp. 247–256.

Kullback, S. (1967). A lower bound for discrimination in terms of variation. *IEEE Trans. Information Theory*, 13, pp. 126-127.

LeCam, L. (1973). "Convergence of estimates under dimensionality restrictions," *Annals of Statistics*, vol. 1, pp. 38–53.

Ledoux, M. (1996). "On Talagrand's deviation inequalities for product measures," *ESAIM: Probability and Statistics*, vol. 1, pp. 63–87.

Liese, F. and Vajda, I. (1987). *Convex statistical distances*. Teubner, Leipzig.

Lugosi, G. and Nobel, A. (1996). "Consistency of data-driven histogram methods for density estimation and classification," *Annals of Statistics*, vol. 24, pp. 687–706.

Macintyre, A. and Sontag, E. D. (1993). "Finiteness results for sigmoidal neural networks," in: *Proceedings of the 25th Annual ACM Symposium on the Theory of Computing*, pp. 325–334, Association of Computing Machinery, New York.

Massart, P. (1990). "The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality," *Annals of Probability*, vol. 18, pp. 1269–1283.

McDiarmid, C. (1989). "On the method of bounded differences," in: *Surveys in Combinatorics 1989*, vol. 141, pp. 148–188, London Mathematical Society Lecture Notes Series, Cambridge University Press, Cambridge.

Nadaraya, E. A. (1974). "On the integral mean square error of some nonparametric estimates for the density function," *Theory of Probability and its Applications*, vol. 19, pp. 133–141.

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

Poggio, T. and Girosi, F. (1990). "A theory of networks for approximation and learning," *Proceedings of the IEEE*, vol. 78, pp. 1481–1497.

Rényi, A. (1959). On the dimension and entropy of probability distributions. *Acta Math. Acad. Sci. Hungar.*, 10, pp. 193-215.

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

Rosenblatt, M. (1971). “Curve estimates,” *Annals of Mathematical Statistics*, vol. 42, pp. 1815–1842.

Sanov, I. N. (1957). On the probability of large deviations of random variables. *Mat. Sb.*, 42, pp. 11-44 (English translation in *Sel. Transl. Math. Statist. Prob.*, 1, (1961), pp.213-244).

Scheffé, H. (1947). A useful convergence theorem for probability distributions. *Ann. Math. Statist.* 18, pp. 434-458.

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

Stone, C. J. (1984). “An asymptotically optimal window selection rule for kernel density estimates,” *Annals of Statistics*, vol. 12, pp. 1285–1297.

Szarek, S. J. (1976). “On the best constants in the Khintchine inequality,” *Studia Mathematica*, vol. 63, pp. 197–208.

Talagrand, M. (1995). “Concentration of measure and isoperimetric inequalities in product spaces,” *Institut des Hautes Etudes Scientifiques. Publications Mathématiques*, vol. 81, pp. 73–205.

Talagrand, M. (1996). “A new look at independence,” *Annals of Probability*, vol. 24, pp. 1–34.

Vajda, I. (1970). Note on discrimination information and variation. *IEEE Trans. Information Theory*, IT-16, 771-773.

Vajda, I. (1989). *Theory of Statistical Inference and Information*. Kluwer Academic Publishers.

Vapnik, V. N. and Chervonenkis, A. Ya. (1971). “On the uniform convergence of relative frequencies of events to their probabilities,” *Theory of Probability and its Applications*, vol. 16, pp. 264–280.

Watson, G. S. and Leadbetter, M. R. (1963). “On the estimation of the probability density, I,” *Annals of Mathematical Statistics*, vol. 34, pp. 480–491.

Wegkamp, M. (2000). “Quasi universal bandwidth selection for kernel density estimators,” *Canadian Journal of Statistics*. To appear.

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

Yatracos, Y. G. (1985). “Rates of convergence of minimum distance estimators and Kolmogorov’s entropy,” *Annals of Statistics*, vol. 13, pp. 768–774.

Yatracos, Y. G. (1988). “A note on $L_1$ consistent estimation,” *Canadian Journal of Statistics*, vol. 16, pp. 283–292.