Prob. Theory and Math. Stat., pp. 000–000
M.L. Puri (Ed.)

INEQUALITIES FOR A NEW DATA-BASED METHOD FOR SELECTING NONPARAMETRIC DENSITY ESTIMATES

LUC DEVROYE,
GÁBOR LUGOSI, AND FREDERIC UDINA
School of Computer Science
McGill University
Montreal, Canada H3A 2A7

Department of Economics and Bussiness
Universitat Pompeu Fabra
Ramon Trias Fargas, 25-27
08005 Barcelona, Spain

ABSTRACT

We develop a general method to select an estimate from any given family of (regular and additive) nonparametric density estimates. We provide explicit non-asymptotic density-free inequalities that relate the L_1 error of the selected estimate with that of the best possible estimate in the family, and study in particular the connection between the richness of the class of density estimates and the performance bound. For example, our method allows one to pick the bandwidth and kernel order in the kernel estimate simultaneously and still assure that for *all densities*, the L_1 error of the corresponding kernel estimate is not larger than about three times the error of the estimate with the optimal smoothing factor and kernel plus a constant times $\sqrt{\log n/n}$, where n is the sample size, and the constant only depends on the complexity of the family of kernels used in the estimate. Among many possible applications we include here multivariate kernel estimates and transformed kernel estimates.

The first author's work was supported by NSERC Grant A3456 and by FCAR Grant 90-ER-0291. The work was supported by DGES Grant PB96-0300.

1. INTRODUCTION

We are given an i.i.d. sample $X_1, \ldots, X_n$ drawn from an unknown density f on $\mathbb{R}^d$. A density estimate $f_n(x) = f_n(x, X_1, \ldots, X_n)$ is a real-valued measurable function of its arguments. Among others, we consider the Akaike-Parzen-Rosenblatt density estimate

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

where $K : \mathbb{R}^d \to \mathbb{R}$ is a fixed kernel with $\int K = 1$, $K_h(x) = (1/h^d) K(x/h)$, and $h > 0$ is the smoothing factor (Akaike, 1954; Parzen, 1962; Rosenblatt, 1956). Many data-dependent choices for h have been proposed in the literature. Most perform well for restricted classes of densities. An exception may be found in the recent work of Devroye and Lugosi (1996, 1997), where data-dependent smoothing factors H are introduced for which

$$\sup_f \limsup_{n \to \infty} \frac{\mathbf{E} \int |f_{nH} - f|}{\inf_h \mathbf{E} \int |f_{nh} - f|} \leq 3 \ ,$$

whenever the kernel K is nonnegative, Lipschitz, and of a compact support. In this paper, we continue the study and propose bandwidths for transformed kernel estimates, variable kernel estimates, and kernel estimates with joint choice of K and h. Explicit non-asymptotic performance guarantees are provided that are uniform over all f. As the same principle may be applied to a host of other estimators, including series estimates, partitioning estimates, various brands of histograms, and tree-based methods, it is advantageous to derive the theory in a general setting (as is done in the next section). To keep the length of the paper reasonable, results on the other methods will be reported elsewhere.

2. THE BASIC ESTIMATE

Let Θ be an abstract set of parameters, and assume that each $\theta \in \Theta$ determines a density estimate $f_{n,\theta}$ for each n. The L_1 error of the estimate $f_{n,\theta}$ is denoted by

$$J_{n,\theta} = \int |f - f_{n,\theta}|.$$

Let $m < n$ (tipically $m \ll n$), and define $\mathcal{A}_\Theta$ as the *Yatracos class* of subsets of $\mathbb{R}^d$ (corresponding to the family of density estimates $f_{n,\theta}$, $\theta \in \Theta$) as the class of all sets of the form

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

We select a parameter θ_n from Θ by minimizing the distance

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

over all $\theta \in \Theta$, where μ_m denotes the empirical measure defined by the subsample $X_{n-m+1}, \ldots, X_n$. The class of parameters may include bandwidths, but also kernels from a class of kernels, parameters in nonlinear transformations, and so forth. There are no a priori restrictions on the size.

A density estimate g_n is called *additive* if it is of the form

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

where $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a measurable function such that for all $y \in \mathbb{R}^d$, $\int_{\mathbb{R}^d} K(x,y)dx = 1$. We say that the additive estimate g_n is *regular* if for each x, $\mathbf{E}K(x, X) < \infty$.

We will use the notion of shatter coefficient as in the work of Vapnik and Chervonenkis (1971):

$$s(\mathcal{A}_\Theta, \ell) = \sup_{y_1, \ldots, y_\ell \in \mathbb{R}^d} |\{\{y_1, \ldots, y_\ell\} \cap A : A \in \mathcal{A}_\Theta\}| \ ,$$

the maximal number of different subsets of a set of ℓ points which can be intersected by sets in $\mathcal{A}_\Theta$. This will be used to measure the richness of classes of density estimates. The first result upon which many of the subsequent results are built is the following non-asymptotic inequality:

THEOREM 1. *Let the set Θ determine a class of regular additive density estimates. Then for all n, $m \le n/2$, Θ, and f,*

$$\mathbf{E}\int |f_{n-m,\theta_n} - f| \le 3 \inf_{\theta\in\Theta} \mathbf{E}\int |f_{n,\theta} - f| \left(1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}}\right) + \\ + \sqrt{\frac{8\log\left(4e^8 s(\mathcal{A}_\Theta, m^2)\right)}{m}} .$$

Note that whevever $s(\mathcal{A}_\Theta, \ell)$ is bounded by a polynomial $n^{k_1}\ell^{k_2}$ of n and ℓ, we have $s(\mathcal{A}_\Theta, m^2) \le n^{k_1} m^{2k_2} \le n^{k_1+2k_2}$, and consequently

$$\sqrt{\frac{8\log\left(4e^8 s(\mathcal{A}_\Theta, m^2)\right)}{m}} = O\left(\sqrt{\frac{\log n}{m}}\right).$$

In the examples below, all bounds for $s(\mathcal{A}_\Theta, \ell)$ will be polynomial in n and ℓ. Furthermore, in this case, if $m \sim n/\log n$, then

$$\mathbf{E}\int |f_{n-m,\theta_n} - f| \le 3 \inf_{\theta\in\Theta} \mathbf{E}\int |f_{n,\theta} - f| \left(1 + O\left(1/\sqrt{\log n}\right)\right) + O\left(\frac{\log n}{\sqrt{n}}\right) .$$

Because in most cases of interest, the optimal L_1 error tends to zero much slower than $1/\sqrt{n}$, this bound essentially says that for polynomial shatter coefficients, we have asymptotically a performance that is guaranteed to be within a factor of 3 of the optimal performance, and this without placing any restrictions on the density f. The proof of Theorem 1 is a minor modification of some arguments appearing in Devroye and Lugosi (1997). The details may be found in the Appendix below.

3. STANDARD KERNEL ESTIMATE: RIEMANN KERNELS

A Borel set A of $\mathbb{R}^d$ is called a star interval if for any $y \in \mathbb{R}^d$, $\{t \in \mathbb{R} : ty \in A\}$ is an interval. Thus, all convex sets are star intervals. A kernel K is said to be Riemann of order k if there exist star intervals $A_1, \ldots, A_k$ and real numbers a_i such that

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

where I_A denotes the indicator function of a set A. We require furthermore that $\int K = 1$. We will call the smallest such k the Riemann order, which should not be confused with the order of a kernel, which is the smallest positive integer s such that $\int x^s K(x)dx \ne 0$, and is in this sense only defined for univariate kernels.

The standard Akaike-Rosenblatt-Parzen kernel estimate is

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

When K is fixed and h is chosen by the method described above (so that $\theta = h$ and $\Theta = \{\theta \in \mathcal{R} : \theta > 0\}$), Theorem 1 applies with the following shatter coefficient:

LEMMA 1 (DEVROYE AND LUGOSI, 1997). *For the kernel estimate with Riemann kernel of order k,*

$$s(\mathcal{A}_\Theta, \ell) \leq (\ell + 1)(1 + 2k\ell(n-m))^2 \leq 18k^2n^2\ell^3 .$$

Let us now widen the scope a bit and pick a Riemann kernel from a finite class of N Riemann kernels, $\mathcal{K} = \{K_1, \ldots, K_N\}$, and choose the bandwidth h simultaneously as well. This is done by formally putting $\Theta = \{(h,j) : h > 0, j \in \{1,\ldots,N\}\}$. Again, Theorem 1 aplies, but now with a slightly larger shatter coefficient:

LEMMA 2. *Consider the class Θ in which $h > 0$ and $K \in \mathcal{K}$ are the free parameters, and assume that all kernels in $\mathcal{K}$ are Riemann of order not exceeding k. Then*

$$s(\mathcal{A}_\Theta, \ell) \leq 18k^2n^2\ell^3N^2 .$$

PROOF. We generalize a proof from Devroye and Lugosi (1997). Set $r = n - m$. We first consider $N = 2$, and let the kernels in $\mathcal{K}$ be K and L, and assume without loss of generality that their Riemann orders are exactly k. Define the vector

$$z_u = \left(\sum_{i=1}^{r} K\left(\frac{y_1 - X_i}{u}\right), \ldots, \sum_{i=1}^{r} K\left(\frac{y_\ell - X_i}{u}\right)\right) \in \mathrm{IR}^\ell.$$

As $u \uparrow \infty$, each component of z_u changes every time $(y_j - X_i)/u$ enters or leaves a set A_l, $1 \leq l \leq k$ for some X_i, $1 \leq i \leq r$, where the A_l's are the star intervals in the definition of K. 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_l, the number of different values a component can take in its history (as $u \uparrow \infty$) is clearly bounded by $2kr$. As there are ℓ components, the cardinality of the set of different values of z_u is bounded by

$$|\{z_u : u > 0\}| \leq 1 + 2k\ell r.$$

If we define z'_u similarly as z_u, but replace K in the definition by L, then we have

$$|\{z'_u : u > 0\}| \leq 1 + 2k\ell r$$

as well. Therefore,

$$|\{(z_u, z'_v) : u, v > 0\}| \leq (1 + 2k\ell r)^2.$$

and the same bound applies for the pairs (z_u, z_v), (z'_u, z'_v) and (z'_u, z_v).
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, $1 \leq i \leq \ell$. Thus,

$$\begin{aligned} \left|\left\{\{y_1, \ldots, y_\ell\} \cap A_{u,v} : (u, v) \in U_{(w,w')}\right\}\right| &\leq \\ \leq \left|\left\{\left(I_{w_1 \geq cw'_1}, \ldots, I_{w_\ell \geq cw'_\ell}\right) : c \geq 0\right\}\right| &\leq \ell + 1. \end{aligned}$$

But then

$$|\{\{y_1, \ldots, y_\ell\} \cap A_{u,v} : (u, v) > 0\}| \leq (\ell + 1)\left|U_{(w,w')}\right| \leq (\ell + 1)(1 + 2k\ell r)^2.$$

The same bound applies for the three other types of pairs, (z_u, z_v), (z'_u, z'_v) and (z'_u, z_v). Thus,

$$s(\mathcal{A}_\Theta, \ell) \leq 4(\ell + 1)(1 + 2k\ell r)^2 \leq 8\ell(3k\ell r)^2 = 72\ell^3 k^2 r^2 \ .$$

If we have a choice between N kernels, we apply the bound not 4 times, but N^2 times, for all possible pairings (with repetition), to obtain

$$s(\mathcal{A}_\Theta, \ell) \leq N^2(\ell + 1)(1 + 2k\ell r)^2 \leq 2N^2\ell(3k\ell r)^2 = 18N^2 m^3 k^2 r^2 \ . \square$$

Lemma 2 permits us to obtain fine inequalities even when the kernel is freely picked from a finite class. However, in all cases, the kernels have to be Riemann of finite order. In the next section, we deal with the joint selection of h and K for general (non-Riemann) K, and this could even include kernels of infinite order.

4. STANDARD KERNEL ESTIMATES: GENERAL KERNELS

If K is not Riemann, we say that it is Riemann approximable if for each n there exists a finite number k such that there exists a Riemann kernel K' of order k with

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

Note that this is always possible if K is Riemann integrable. The smallest such k will be called the *kernel complexity* κ_n. If there is a finite class of kernels $K \in \mathcal{K}$, then we need to find Riemann approximations K' for each K individually. A kernel estimate with Riemann kernel K' is piecewise constant and thus easy to work with in simulations.

Define the kernel estimates

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

for all $h > 0$ and $K \in \mathcal{K}$. Let the pair (H, K) (where K has Riemann approximation K') be selected from $\Theta = (0, \infty) \times \mathcal{K}$ such that

$$\sup_{A \in \mathcal{A}} \left| \int_A f_{n-m,K',h} - \mu_m(A) \right|$$

is minimal where $\mathcal{A}$ is defined as the collection of all sets

$$\{x : f_{n-m,L',u}(x) \ge f_{n-m,M',v}(x)\}$$

with $u, v > 0$ and L', M' are Riemann approximations of kernels L, M from $\mathcal{K}$. After the selection, the Riemann kernels are no longer needed. Finally, our estimate is $f_{n-m,K,H}$. We may also use $f_n = f_{n,K,H}$ and refer to Devroye and Lugosi (1996) for analysis of this situation. Sánchez-Sellero and de Uña (Devroye, 1997) report good experimental results if all data are used and not just the first $n - m$ data points. For a practical implementation and experimental comparison, we refer to Devroye (1997). Finally, one may wonder if the derivation via Riemann kernels is really needed. It seems that the combinatorial arguments that will follow may be made to work for certain classes of kernels such as polynomials, but in any case, the generality achieved here will be lost.

We offer the following non-asymptotic bound:

THEOREM 2. *Consider the kernel density estimate with joint choice of H and K as described above, where K is taken from a class $\mathcal{K}$ of N kernels with kernel complexities uniformly bounded by κ_n. Then for all n, $m \le n/2$, d, and f,*

$$\begin{aligned}\mathbf{E}\int |f_{n-m,K,H} - f| \le\; & 3\inf_{h>0,L\in\mathcal{K}} \mathbf{E}\textstyle\int |f_{n,L,h} - f| \left(1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}}\right) \\ & + \sqrt{8\log\left(72e^8N^2\kappa_n^2n^8\right)/m} + 27/n \;.\end{aligned}$$

For n even and $m = n/2$, we thus have

$$\begin{aligned}\mathbf{E}\int |f_{n-m,K,H} - f| \le\; & \inf_{h>0,L\in\mathcal{K}} \mathbf{E}\textstyle\int |f_{n,L,h} - f| \left(9 + 24/\sqrt{2}\right) \\ & + \sqrt{16\log\left(72e^8N^2\kappa_n^2n^8\right)/n} + 27/n \;.\end{aligned}$$

PROOF. Note that

$$\mathbf{E}\int |f_{n-m,K,H} - f| \le \mathbf{E}\int |f_{n-m,K',H} - f| + \frac{1}{n} \;.$$

Furthermore,

$$\inf_{h>0,L\in\mathcal{K}} \mathbf{E}\int |f_{n,L',h} - f| \le \inf_{h>0,L\in\mathcal{K}} \mathbf{E}\int |f_{n,L,h} - f| + \frac{1}{n} \;.$$

Thus, a combination with Theorem 1 then yields, with the appropriate definition of Θ, and writing $B_{m,n}$ for $\left(1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}}\right)$,

$$\begin{aligned}& \mathbf{E}\int |f_{n-m,K,H} - f| \le \mathbf{E}\int |f_{n-m,K',H} - f| + \frac{1}{n} \le \\ & \le\; 3\inf_{h>0,L\in\mathcal{K}} \mathbf{E}\int |f_{n,L',h} - f|B_{m,n} + \sqrt{8\log\left(4e^8s(\mathcal{A}_\Theta, m^2)\right)/m} + 1/n \\ & \le\; 3\inf_{h>0,L\in\mathcal{K}} \mathbf{E}\int |f_{n,L,h} - f|B_{m,n} + \sqrt{8\log\left(4e^8s(\mathcal{A}_\Theta, m^2)\right)/m} + 27/n \\ & \le\; 3\inf_{h>0,L\in\mathcal{K}} \mathbf{E}\int |f_{n,L,h} - f|B_{m,n} + \sqrt{8\log\left(4e^818N^2m^6\kappa_n^2n^2\right)/m} + 27/n \\ & \le\; 3\inf_{h>0,L\in\mathcal{K}} \mathbf{E}\int |f_{n,L,h} - f|B_{m,n} + \sqrt{8\log\left(72e^8N^2\kappa_n^2n^8\right)/m} + 27/n \;,\end{aligned}$$

where we used Lemma 2. □

The above inequality is useful when N is finite. The other quantity of interest is κ_n. We briefly recall some bounds from Devroye and Lugosi (1997).

Kernels with $\kappa_n = O(n^b)$ for some finite b are said to be *polynomially Riemann approximable*. All kernels of practical interest are in this class, as we will see below. For fixed N, the last two terms in the upper bound of Theorem 2 are then $O(\sqrt{\log n/m})$, just as in the case of Riemann kernels. Obviously, if K is Riemann of order k, then $\kappa_n \le k$. Symmetric unimodal kernels on the real line have $\kappa_n \le 8nK(0)\beta + 10$, where β is the last positive value for which $\int_\beta^\infty K \le 1/(4n)$. If $K(x) \le aI_{[-b,b]}(x)$ and K is symmetric, nonnegative, and unimodal (such as the Epanechnikov-Bartlett kernel), then $\kappa_n \le 8nab + 10$. For the normal density, we have $\kappa_n \le \frac{8n\sqrt{\log n}}{\sqrt{\pi}} + 10$. Products of polynomially approximable kernels in $\mathbb{R}^d$ and multivariate normal densities are also polynomially Riemann approximable.

We finish this section by noting the impact of Theorem 2 if all N kernels are of orders up to and including an even number s, that is, each of the kernels K is bounded, symmetric, and has finite nonzero s-th moment and at least one kernel has zero i-th moments for $0 < i < s$. Then regardless of the density and the choice of h,

$$\liminf_{n\to\infty} n^{s/(2s+1)} \inf_h \mathbf{E} \int |f_{n,K,h} - f| > 0$$

(Devroye, 1988, page 1173). For such higher order kernels, let $m = o(n)$ such that $m/(n^{2s/(2s+1)} \log n) \to \infty$. Then if $\kappa_n = O(n^\alpha)$ for some finite α, uniformly over the N kernels,

$$\mathbf{E} \int |f_{n,K,H} - f| \le (3 + o(1)) \inf_{h, L\in\mathcal{K}} \mathbf{E} \int |f_{n,L,h} - f| + o(n^{-s/(2s+1)}) \ ,$$

and therefore

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

Thus, Theorem 2 shows asymptotic optimality to within a factor of 3 for all (finite collections of) kernels of finite order.

5. MULTIPARAMETER KERNEL ESTIMATES—PRODUCT KERNELS

Consider the kernel estimate

$$f_{n,\theta}(x) = \frac{1}{n} \sum_{i=1}^{n} K_\theta(x - X_i),$$

where $\theta = (h_1, \ldots, h_d)$ is a vector of positive smoothing factors, and

$$K_\theta(x) = \prod_{j=1}^{d} \frac{1}{h_j} K_j\left(\frac{x^{(j)}}{h_j}\right),$$

where $K_1, \ldots, K_d$ are fixed one-dimensional kernels integrating to one, and $x^{(j)}$ is the j-th component of x. Thus, we let the smoothing factor vary in each direction. The issue here is the data-based choice of the smoothing factors. For brevity, we consider only the simplest possible kernels. The bound of Theorem 1 is applicable if we can compute the shatter coefficient.

LEMMA 3. *Assume that for each j, $K_j = I_{A_j}(x)$, where $A_j = [a_j, a_j+1]$ is an interval of unit length. Then for $\ell n \geq 2d$ we have*

$$s(\mathcal{A}_\Theta, \ell) \leq (\ell+1)\left(\frac{\ell n e}{2d}\right)^{4d}.$$

Lemma 3 shows that the shatter coefficient is polynomial in n and ℓ. Therefore, the same bounds apply as for the univariate or single h kernel estimates, with just a different coefficient in the additive term of the bound. It is quite remarkable that adjusting d parameters is not appreciably more difficult than adjusting one parameter. For general products of Riemann kernels, bounds similar to those of Lemma 3 may be obtained. For products of polynomially Riemann approximable kernels, one needs to optimize a criterion that involves the Riemann approximations, just as in the previous two sections. The details are omitted.

PROOF OF LEMMA 3. Denote the j-th component of y_t by $y_t^{(j)}$, $t \leq \ell$, $j \leq d$, and the j-th component of X_i by $X_i^{(j)}$, $i \leq n-m$, $j \leq d$. For each $\theta \in \Theta$ define the vector

$$z_\theta = \left(z_\theta^{(1)}, \ldots, z_\theta^{(\ell)}\right) =$$
$$\left(\sum_{i=1}^{n-m}\prod_{j=1}^{d} K_j\left(\frac{y_1^{(j)} - X_i^{(j)}}{h_j}\right), \ldots, \sum_{i=1}^{n-m}\prod_{j=1}^{d} K_j\left(\frac{y_\ell^{(j)} - X_i^{(j)}}{h_j}\right)\right).$$

Observe that for each $t \leq \ell$ and $i \leq n-m$

$$\prod_{j=1}^{d} K_j\left(\frac{y_t^{(j)} - X_i^{(j)}}{h_j}\right) = 1 \qquad \text{if and only if} \quad y_t - X_i \in R_\theta,$$

where R_θ denotes the rectangle $[a_1/h_1, (a_1+1)/h_1] \times \cdots \times [a_d/h_d, (a_d+1)/h_d]$. Since there are $\ell(n-m)$ possible values for $y_t - X_i$, the number of different

values the vector z_θ can take as θ varies through Θ is at most $s(\mathcal{B}, \ell(n-m))$, where $\mathcal{B}$ is the class of all rectangles in $\mathbb{R}^d$. But it is well-known (see, e.g., Devroye, Györfi, and Lugosi (1996, p.220) that for $\ell(n-m) \geq 2d$ the shatter coefficients of this class are bounded as $s(\mathcal{B}, \ell(n-m)) \leq \left(\frac{\ell(n-m)e}{2d}\right)^{2d}$.
It follows that

$$|\{(z_{\theta_1}, z_{\theta_2}) : \theta_1, \theta_2 \in \Theta\}| \leq \left(\frac{\ell(n-m)e}{2d}\right)^{4d}.$$

The rest of the proof is now standard: Let

$$\mathcal{W} = \{(w, w') : (w, w') = (z_{\theta_1}, z_{\theta_2}) \text{ for some } \theta_1, \theta_2 \in \Theta\}.$$

For fixed $(w, w') \in \mathcal{W}$, let $U_{(w,w')}$ denote the collection of all (θ_1, θ_2) such that $(z_{\theta_1}, z_{\theta_2}) = (w, w')$. For $(\theta_1, \theta_2) \in U_{(w,w')}$, we have

$$y_t \in A_{\theta_1,\theta_2} \quad \text{if and only if} \quad z_{\theta_1}^{(t)} \prod_{j=1}^{d} \frac{1}{h_{j,1}} \geq z_{\theta_2}^{(t)} \prod_{j=1}^{d} \frac{1}{h_{j,2}},$$

where $\theta_i = (h_{1,i}, \ldots, h_{d,i})$ for $i = 1, 2$. Within the set $U_{(w,w')}$, $z_{\theta_1}^{(t)}$ and $z_{\theta_2}^{(t)}$ are fixed for all t, and therefore

$$\begin{aligned} \left|\left\{\{y_1, \ldots, y_\ell\} \cap A_{u,v} : (u,v) \in U_{(w,w')}\right\}\right| &\leq \\ \leq \left|\left\{\left(I_{w_1 \geq c w_1'}, \ldots, I_{w_\ell \geq c w_\ell'}\right) : c \geq 0\right\}\right| &\leq \ell + 1, \end{aligned}$$

where $w_1, \ldots, w_\ell$ and $w_1', \ldots, w_\ell'$ denote the components of the vectors w and w', respectively. But then

$$|\{\{y_1, \ldots, y_\ell\} \cap A_{u,v} : (u,v) > 0\}| \leq (\ell+1) \left|U_{(w,w')}\right| \leq (\ell+1) \left(\frac{\ell(n-m)e}{2d}\right)^{4d}.$$

□

6. MULTIPARAMETER KERNEL ESTIMATES—ELLIPSOIDAL KERNELS

Next we consider the kernel estimate

$$f_{n,\theta}(x) = \frac{1}{n} \sum_{i=1}^{n} K_\theta(x - X_i),$$

where $\theta = \Sigma$, and Σ is a positive definite symmetric $d \times d$ matrix, and

$$K_\theta(x) = v_\theta I_{\{x^T\Sigma^{-1}x \le 1\}}.$$

Here v_θ is a normalizing factor such that $\int K_\theta = 1$, and x^T denotes the transpose of the vector x. In this case, for $\ell(n-m) \ge d^2 + d + 2$, we have

$$s(\mathcal{A}_\Theta, \ell) \le (\ell+1)\left(\frac{\ell(n-m)e}{d^2/2 + d/2 + 1}\right)^{d^2+d+2}.$$

The proof is exactly the same as for the case of product kernels with the only difference that the shatter coefficients of the class $\mathcal{E}$ of ellipsoids (i.e., class of sets of the form $E_\theta = \{x : x^T\Sigma^{-1}x \le 1\}$) is bounded by

$$s(\mathcal{E}, \ell(n-m)) \le \left(\frac{\ell(n-m)e}{d^2/2 + d/2 + 1}\right)^{d^2/2+d/2+1}.$$

whenever $\ell(n-m) \ge d^2 + d + 2$ (since the VC dimension of $\mathcal{E}$ is bounded by $d^2/2+d/2+1$, see, e.g., Devroye, Györfi, and Lugosi (1996, p.221). Although it is computationally challenging to optimize all $\binom{d}{2}$ entries in a matrix, at least in theory, we can set up a method (by picking m) such that asymptotically, the performance is about three times or less times the best possible performance over all such matrices. Again, no conditions are placed on the density or the values of the entries in the matrix.

Similarly to the univariate case, the argument may be extended via Riemann approximations to the class of estimates with $K_\theta(x) = v_\theta L(x^T\sigma^{-1}x)$, where $L : \mathbb{R}^+ \to \mathbb{R}$ is a fixed function. The details are omitted.

7. THE TRANSFORMED KERNEL ESTIMATE

The transformed kernel estimate on the real line was introduced by Devroye and Györfi (1985) in an attempt to reduce the L_1 error in a relatively cheap manner. The data are transformed by a smooth monotone transform $y = T(x)$, the transformed density is estimated by the kernel estimate, and the estimate is then subjected to the inverse transformation. As this leaves the L_1 error unaltered, it suffices to study the L_1 error in the transformed space, and hence the interest of such estimates. In particular, it is known that heavy tails are to be avoided for kernel estimates (Devroye and Györfi, 1985). Thus, transforms that compact and compress the data are called for. Ideally, the transformed density should be triangular. Thus, we consider the joint optimization over (h, a), where h is the smoothing factor, and a is a parameter of the transformation. For simplicity, we will consider the Box-Cox transformations, with which statisticians and data analysts are familiar. We will show that we can jointly pick h and a in an asymptotically optimal manner, still modulo a factor 3, without placing any restrictions on the density or the parameters. The transformations considered here are only useful to treat tail problems. A similar analysis may be carried out for piecewise linear transformations, the transformation being restricted to consist of a fixed number of segments, but otherwise arbitrary. Such estimators are close in spirit to variable kernel estimators. For practical data-based versions of other transformations, we refer to Wand, Marron and Ruppert (1991) and Ruppert and Cline (1994).

In general, the transformed kernel estimate is

$$f_{n,T}(x) = \frac{1}{n} \sum_{i=1}^{n} K(T(x) - T(X_i))T'(x),$$

where K is a kernel with $\int K = 1$, and $T : \mathbb{R} \to \mathbb{R}$ is a strictly monotonically increasing almost everywhere differentiable transformation. Clearly, $\int f_{n,T} = 1$. If $T = ax + b$ is linear, then $f_{n,T}$ is just the ordinary kernel estimate with smoothing factor $h = 1/a$. Here we are concerned with the data-based choice of T. Clearly, the collection of possible transformations has to be restricted somehow. Among the many possiblitities, we only consider two representative examples.

Box-Cox transformations. Consider now the family $\{T_a : a \in [0,1]\}$ of transformations defined, for $x > 0$, by

$$T_a(x) = \begin{cases} \frac{x^a - 1}{a} & \text{if } a > 0 \\ \log x & \text{if } a = 0. \end{cases}$$

These functions are often used to transform the (nonnegative) data so that large tails become more manageable. We consider kernel estimates defined on the transformed data. In particular, we study the joint data-based selection of the tranformation (i.e., the value of a) and the bandwidth. For simplicity, we again only consider the naive kernel $K = I_{[-1/2,1/2]}$. Therefore, the class of estimates $\{f_{n,\theta} : \theta \in \Theta\}$ is defined by

$$f_{n,\theta}(x) = \frac{1}{nh} \sum_{i=1}^{n} I_{\{|T_a(x) - T_a(X_i)| \le h/2\}} x^{a-1},$$

where $\theta = (a, h)$ and $\Theta = [0,1] \times (0, \infty)$. Note that we assume that all data points are positive and $f_{n,\theta}(x)$ is only defined for $x > 0$. Again, to see if the proposed parameter selection method works, it suffices to bound $s(\mathcal{A}_\Theta, \ell)$.

LEMMA 4. *Let $\mathcal{A}_\Theta$ denote the Yatracos class corresponding to the family of kernel estimates on $\mathbb{R}^+$ based on all Box-Cox transformations T_a, $a \in [0,1]$ and all smoothing factors $h > 0$. If $\ell \geq 2$ and $n - m \geq 2$, then*

$$s(\mathcal{A}_\Theta, \ell) \leq \frac{9}{4}\ell^6 (n-m)^4.$$

PROOF. In the proof we use a simple lemma which is an easy modification of Lemma 25.2 of Devroye, Györfi, and Lugosi (1996):

LEMMA 5. *If $b_1, \ldots, b_k, c_1, \ldots, c_k \in \mathbb{R}$, then the function*

$$g(x) = \sum_{i=1}^{k} b_i e^{c_i x}$$

is either identically zero or takes the value 0 for at most $k-1$ different places.

LEMMA 6 (COVER (1965)). *Let $\mathcal{A}$ be the class of sets of the form $\{x : a^T x \geq b\} \subset \mathbb{R}^d$, where $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are arbitrary. Then*

$$s(\mathcal{A}, n) \leq 2\sum_{i=0}^{d} \binom{n-1}{i} \leq 2(n-1)^d + 2 \leq 2n^d.$$

PROOF OF LEMMA 4. Consider the $\ell \times (n-m)$ matrix $z_{a,h}$ with entries

$$z_{a,h}^{t,i} = I_{\{|X_i^a - y_t^a| < ah\}} \quad t = 1, \ldots, \ell, \; i = 1, \ldots, n-m.$$

First we bound the number of possible different values of the matrix $z_{a,h}$ as $(a,h) \in [0,1] \times (0,\infty)$. Observe that in the set $[0,1] \times (0,\infty)$, each pair (t,i) defines a curve given by

$$u^a - v^a - ah = 0 \qquad \text{where} \quad u = \max(X_i, y_t) \text{ and } \quad v = \min(X_i, y_t).$$

If two curves $u^a - v^a - ah = 0$ and $w^a - z^a - ah = 0$ intersect at the point (a,h), then

$$e^{a \log u} - e^{a \log v} - e^{a \log w} + e^{a \log z} = 0.$$

According to Lemma 5, this cannot happen at more than three points (unless $u = w$ and $v = z$). Next we argue that these curves partition the set $[0,1] \times (0,\infty)$ into at most $(3\ell^2(n-m)^2 - \ell(n-m) + 4)/2 \leq (3/2)\ell^2(n-m)^2$ connected regions. This may be easily seen by induction, since if s_N denotes the number of connected regions defined by N such curves, then it is clear that $S_1 = 2$ and $S_{N+1} \leq S_N + 3N + 1$, since any two curves intersect at at most three points. The solution of this recursion is $S_N = (3N^2 - N + 4)/2$. Since inside each region the value of the matrix $z_{a,h}$ is a constant, $(3/2)\ell^2(n-m)^2$ is an upper bound on the number of possible values of the matrix. Therefore,

$$|\{(z_{a,h}, z_{a',h'}) : a, a' \in [0,1], h, h' > 0\}| \leq \frac{9}{4}\ell^4(n-m)^4.$$

Consider now a region in $([0,1] \times (0,\infty))^2$ over which $(z_{a,h}, z_{a',h'})$ is constant, with value, say, (w, w'), and denote the set of such quadruples (a, h, a', h') by $U_{(w,w')}$. Denoting $\theta = (a, h)$ and $\theta' = (a', h')$,

$$y_i \in A_{\theta,\theta'} \qquad \text{if and only if} \qquad \frac{y_t^{a-1}}{h}\sum_{i=1}^{n-m} w^{(t,i)} \geq \frac{y_t^{a'-1}}{h'}\sum_{i=1}^{n-m} {w'}^{(t,i)},$$

or equivalently, if and only if

$$\begin{aligned}(a-1)\log y_t - \log h + \log\left(\sum_{i=1}^{n-m} w^{(t,i)}\right) &\geq \\ \geq (a'-1)\log y_t - \log h' + \log\left(\sum_{i=1}^{n-m} {w'}^{(t,i)}\right)&.\end{aligned}$$

Therefore, denoting $W_t = \sum_{i=1}^{n-m} w^{(t,i)}$ and $W'_t = \sum_{i=1}^{n-m} {w'}^{(t,i)}$, the maximal number of different values of

$$\left(I_{A_{\theta_1,\theta_2}}(y_1), \ldots, I_{A_{\theta_1,\theta_2}}(y_\ell)\right)$$

is at most the number of different values of the vector

$$\left(I_{(a-a')\log y_1 - \log(h/h') + \log(W_1/W'_1) \geq 0} \cdots, I_{(a-a')\log y_t - \log(h/h') + \log(W_t/W'_t) \geq 0}\right)$$

takes as a, a', h, h' all vary through $\mathbb{R}^+$. But this is not more than the maximal number of different ways of dichotomizing ℓ points by 2-dimensional hyperplanes, which, by Lemma 6, is at most ℓ^2 (since $\ell \geq 2$). Collecting bounds, the proof is finished. $\square$

Having Lemma 4, Theorem 1 yields the following bound:

THEOREM 3. *Assume that the basic estimate of Section 2 is used to simultaneously select the Box-Cox transformation T_a, $a \in [0,1]$ and the smoothing factor $h > 0$ for the transformed kernel estimate*

$$f_{n,a,h}(x) = \frac{1}{nh} \sum_{i=1}^{n} I_{\{|T_a(x)-T_a(X_i)| \le h/2\}} x^{a-1}.$$

If f_n denotes the obtained density estimate, then for all $n, m \le n/2$, and each density f over $\mathbb{R}^+$,

$$\begin{aligned} \mathbf{E} \int |f_n - f| \quad &\le \quad 3 \inf_{a\in[0,1],h>0} \mathbf{E} \int |f_{n,a,h} - f| \left(1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}}\right) \\ &\qquad + \sqrt{\frac{8 \log\left(9e^8 m^{12}(n-m)^4\right)}{m}} \; . \end{aligned}$$

For example, if n is even and we take $m = n/2$,

$$\mathbf{E} \int |f_n - f| \le 26 \inf_{a\in[0,1],h>0} \mathbf{E} \int |f_{n,a,h} - f| + 16\sqrt{\frac{\log n}{n}} \; .$$

8. MONTE CARLO SIMULATIONS

For testing the behavior of the proposed parameter selectors we have conducted a series of Monte Carlo simulations. We describe here its graphical and numerical results.

Example 1: Pareto density with transformed kernel

First we consider as target a Pareto density and we use the Box-Cox family of transformations as described in Section 7. To avoid the problem of selecting among infinite sets of parameters, we take only some values of the parameter $a \in [0,1]$ and some bandwidth values.

Consider the Pareto $(1,1)$ density $f(x) = 1/x^2$, $x > 1$, and let the sample size be $N = 1000$. The number of samples is $B = 1000$. For every sample, we select among three values of the Box-Cox parameter (0, 0.35, 0.70) and three bandwidths 0.010, 0.084, and 0.700 in geometric sequence. This gives a total of 9 estimators, and we select the one that minimizes the distance described in Section 2 for the given sample and the nine-element parameter set. 539 times out of 1000 the selected estimator was the *right* one, the one that minimized the IAE (integrated absolute error, or L_1 norm) for the sample. A summary of results is given in Table 1. The average error IAE_s - IAE_b is quite large due to the use of very few bandwidth values, but note that the average relative error (average over samples of $(\mathrm{IAE}_s - \mathrm{IAE}_b)/\mathrm{IAE}_b$, where s subscript denotes the selected estimator and b the *best*) is small.

	Example 1	Example 2
Average error	0.196	.0307
Average relative error	.0563	0.097
Worst relative error	.7716	.7746
prob(rel. error > 10%)	0.218	0.328
prob(rel. error > 50%)	0.002	0.018

Table 1.

Performance of transformed kernel estimators selection for examples 1 and 2. Relative error is $(\mathrm{IAE}_s - \mathrm{IAE}_b)/\mathrm{IAE}_b$ as described in the text.

In figure 1 we show six of the nine involved estimators in a particular sample run (to simplify the figure we don't show the ones with bandwidth 0.01). As expected, it can be seen that bigger bandwidths yield better performance in the tails but are worse in the infinite peak. This shows the convenience of calibrating the transformation to use.

Example 2: Transformed triangular densities

To see how our method can detect the *right* transformation to use, we take a triangular density

$$t(x) = (1 - |x-1|)_+$$

and we back-transform it using the inverse of a Box-Cox(β) transformation. So, we consider as target density

$$f_\beta(x) = x^{(\beta-1)} t\left(\frac{x^\beta - 1}{\beta}\right).$$

If this density is transformed by a Box-Cox(α) transformation, the resulting density will be triangular when $\alpha = \beta$. Given that (see Wand-Devroye(1993)) the triangular density is the easiest to estimate using the Epanechnikov kernel, a good selector might choose β when given several possibilities.

We take $\beta = 0.5$ and we consider five values for the Box-Cox parameter (0, 0.250, 0.500, 0.750, 1.000) and seven bandwidth values from 0.01 to 2 in geometric steps. Note that $\alpha = 1$ gives no transformation at all. For $N = 15000$ we obtained 1000 simulation samples and selected a parameter pair as before. The third column in Table 1 summarizes the L_1 error performance. In Table 2 we collect for each parameter pair the number of times it was the optimal one and the number of times it was selected when at the same time that pair was the optimal one. The figures show that the selector picked up the right pair in 23% of the runs, 52% of them corresponding to the theoretically best transformation.

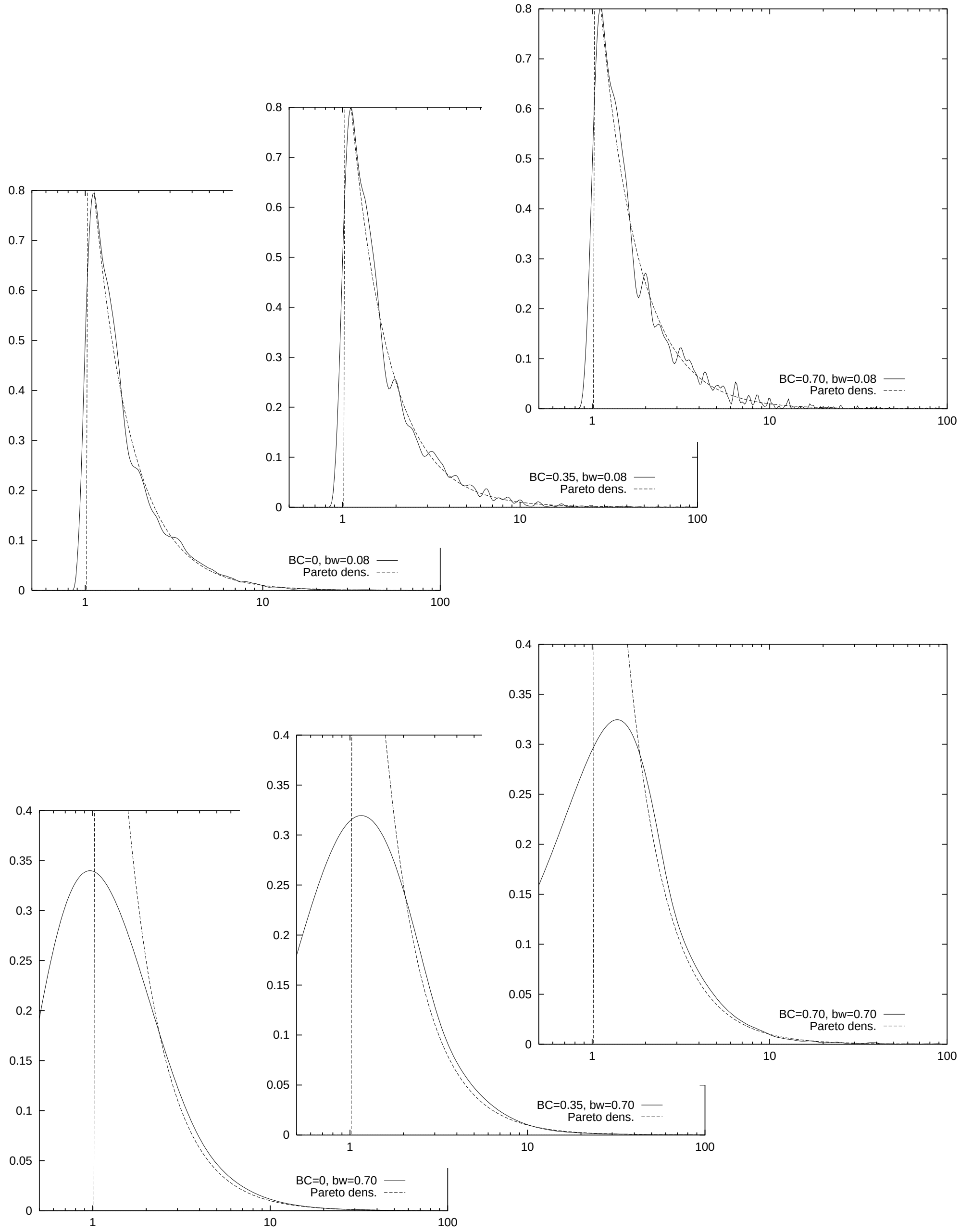

Figure 1. Some of the estimators in competition shown for a particular sample. Upper part is for bandwidth 0.08, lower part for 0.7. In each figure, three Box-Cox parameter values are used: 0, 0.35, and 0.7. Note that horizontal scale is logarithmic and that left tail of the estimators in the lower graphs is not shown.

Bandwidth	Box-Cox parameter				
	0	0.250	0.500	0.750	1.00
0.010	0/0	0/0	0/0	0/0	0/0
0.024	0/0	0/0	0/0	0/0	0/0
0.058	26/6	2/1	0/0	0/0	0/0
0.141	27/6	232/59	476/120	200/30	26/5
0.342	0/0	0/0	0/0	0/0	11/2
0.827	0/0	0/0	0/0	0/0	0/0
2.000	0/0	0/0	0/0	0/0	0/0

Table 2.

In the back-transformed triangular density example, number of times (out of 1000) each parameter pair was the optimal one and the number of these times it was selected.

Note that in this example the sample size must be large because the differences between the estimators involved are very small, as can be appreciated in Figure 2 where we show the five densities resulting from applying the Box-Cox transformation (parameters as above) to f_β, $\beta = 0.5$.

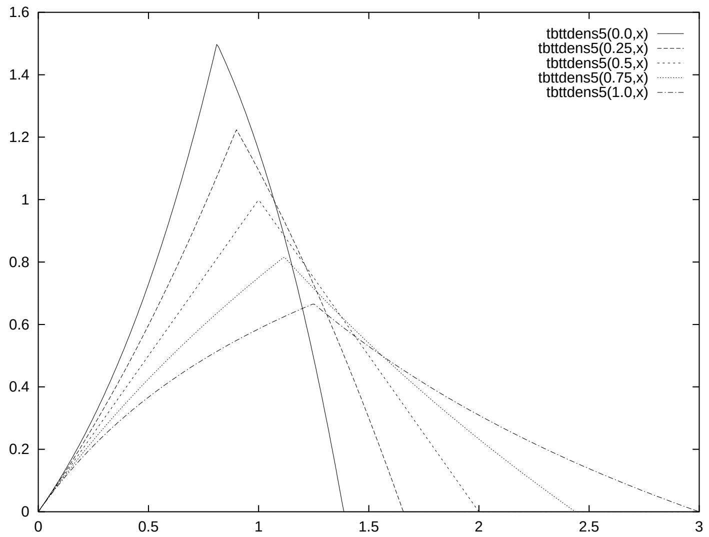

Figure 2. Density f_β ($\beta = 0.5$) transformed by Box-Cox transformations with parameters $\alpha = 0$, 0.250, 0.500, 0.750, 1.000. See Example 2.

A. PROOF OF THEOREM 1

Lemma 7.

$$\int |f_{n-m,\theta_n} - f| \le 3 \inf_{\theta\in\Theta} \int |f_{n-m,\theta} - f| + 4 \sup_{A\in\mathcal{A}_\Theta} \left| \int_A f - \mu_m(A) \right|.$$

Proof. Fix an $\epsilon > 0$, and let $\bar{f}$ be an estimate $f_{n-m,\theta}$ such that

$$\int |\bar{f} - f| \le \int |f_{n-m,\theta} - f| + \epsilon$$

for all $\theta \in \Theta$. Then

$$\begin{aligned}
&\int |f_{n-m,\theta_n} - f| \\
&\le \int |\bar{f} - f| + \int |f_{n-m,\theta_n} - \bar{f}| \\
&= \int |\bar{f} - f| + 2 \sup_{A\in\mathcal{A}_\Theta} \left| \int_A f_{n-m,\theta_n} - \int_A \bar{f} \right| \qquad \text{(by Scheffé's theorem),} \\
&\le \int |\bar{f} - f| + 2 \sup_{A\in\mathcal{A}_\Theta} \left| \int_A f_{n-m,\theta_n} - \mu_m(A) \right| + 2 \sup_{A\in\mathcal{A}_\Theta} \left| \mu_m(A) - \int_A \bar{f} \right| \\
&\qquad \text{(by the triangle inequality)} \\
&\le \int |\bar{f} - f| + 4 \sup_{A\in\mathcal{A}_\Theta} \left| \mu_m(A) - \int_A \bar{f} \right| \qquad \text{(by the definition of } \theta_n\text{)} \\
&\le \int |\bar{f} - f| + 4 \sup_{A\in\mathcal{A}_\Theta} \left| \int_A f - \int_A \bar{f} \right| + 4 \sup_{A\in\mathcal{A}_\Theta} \left| \mu_m(A) - \int_A f \right| \\
&\qquad \text{(by the triangle inequality)} \\
&\le 3 \int |\bar{f} - f| + 4 \sup_{A\in\mathcal{A}_\Theta} \left| \mu_m(A) - \int_A f \right| \qquad \text{(by Scheffé's theorem),} \\
&\le 3 \inf_{\theta\in\Theta} \int |f_{n-m,\theta} - f| + 3\epsilon + 4 \sup_{A\in\mathcal{A}_\Theta} \left| \int_A f - \mu_m(A) \right|.
\end{aligned}$$

Since ϵ is arbitrary, Lemma 7 is proved. □

The following simple lemmas are used in the proof:

LEMMA 8. *Let X and Y be independent random variables, and let $\mathbf{E}Y = 0$. Then $\mathbf{E}|X+Y| \geq \mathbf{E}|X|$.*

LEMMA 9 (DEVROYE AND GYÖRFI, 1985, PAGE 137). *Let $Y_1, \dots, Y_n$ be i.i.d. zero mean random variables. Then*

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

LEMMA 10. *Let Θ be a class of parameters, and assume that each density estimate $f_{n,\theta}$ is additive and regular. If $m > 0$ is a positive integer such that $2m \leq n$, then*

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

PROOF. The proof uses additivity in an essential manner, but otherwise follows the outlines of Devroye and Lugosi (1997). Note the following:

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

The supremum is rewritten as follows:

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

where we used a simple bound from page 23 of Devroye and Györfi (1985). Fix x and θ for now. Introduce

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

and denote the partial sums of Y_i's by $S_j = Y_1 + \cdots + Y_j$. By assumption, for fixed x and θ, the first absolute moment of Y_1 exists. 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

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

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

$$\begin{aligned}
\frac{\mathbf{E}|f_{n-m,\theta}-f_{n,\theta}|}{\mathbf{E}|f_{n,\theta}-\mathbf{E}f_{n,\theta}|} &\le \frac{\frac{m}{n-m}\mathbf{E}|S_{n-m}| + \mathbf{E}|S_m|}{\mathbf{E}|S_n|} \\
&\le \frac{m}{n-m} + \frac{\mathbf{E}|S_m|}{\mathbf{E}|S_n|} \quad (\text{because } \mathbf{E}|S_n| \ge \mathbf{E}|S_{n-m}|) \\
&\le \frac{m}{n-m} + \frac{\mathbf{E}|S_m|}{\sqrt{\frac{\lfloor n/m \rfloor}{8}}\mathbf{E}|S_m|} \quad (\text{by Lemmas 8 and 9}) \\
&\le \frac{m}{n-m} + 4\sqrt{\frac{m}{n}} \quad (\text{if } 2m \le n) \ .
\end{aligned}$$

This implies that for any fixed θ,

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

The lemma now follows without work. □

We now note the following: a variant of the Vapnik-Chervonenkis inequality (Vapnik and Chervonenkis (1971); see Devroye (1982)) states that for $\epsilon > 0$,

$$\mathbf{P}\left\{\sup_{A\in\mathcal{A}_\Theta}\left|\mu_m(A) - \int_A f\right| > \epsilon \middle| X_1,\ldots,X_{n-m}\right\} \le 4e^8 s(\mathcal{A}_\Theta, m^2)e^{-2m\epsilon^2}.$$

This implies by standard bounding that

$$\mathbf{E}\left\{\sup_{A\in\mathcal{A}_\Theta}\left|\mu_m(A) - \int_A f\right| \middle| X_1,\ldots,X_{n-m}\right\} \le \sqrt{\frac{\log\left(4e^8 s(\mathcal{A}_\Theta, m^2)\right)}{2m}}$$

(see Devroye, Györfi, and Lugosi (1996, page 208)). Theorem 1 now follows from this estimate, Lemma 7 and Lemma 10.

REFERENCES

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

T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition", *IEEE Transactions on Electronic Computers*, vol. 14, pp. 326–334, 1965.

K. B. Davis, "Mean integrated square error properties of density estimates", *Annals of Statistics*, vol. 5, pp. 530–535, 1977.

L. Devroye, "Bounds for the uniform deviation of empirical measures", *Journal of Multivariate Analysis*, vol. 12, pp. 72–79, 1982.

L. Devroye, "Asymptotic performance bounds for the kernel estimate", *Annals of Statistics*, vol. 16, pp. 1162–1179, 1988.

L. Devroye, "A note on the usefulness of superkernels in density estimation", *Annals of Statistics*, vol. 20, pp. 2037–2056, 1992.

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

L. Devroye, L. Györfi, *Nonparametric density estimation: the L_1 view* John Wiley, New York, 1985.

L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New York, 1996.

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

L. Devroye and G. Lugosi, "Non-asymptotic universal smoothing factors, kernel complexity and Yatracos classes", *Annals of Statistics*, vol. 25, pp. 2626–2637, 1997.

P. Hall and J. S. Marron, "Choice of kernel order in density estimation", *Annals of Statistics*, vol. 16, pp. 161–173, 1988.

I. A. Ibragimov and R. Z. Khasminskii, "Estimation of distribution density belonging to a class of entire function", *Theory of Probability and its Applications*, vol. 27, pp. 551–562, 1982.

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

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

D. Ruppert and D. B. H. Cline, "Bias reduction in kernel density estimation by smoothed empirical transformations", *Annals of Statistics*, vol. 22, pp. 185–210, 1994.

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

M. P. Wand, L. Devroye, "How easy is a given density to estimate?", *Computational Statistics and Data Analysis*, vol. 16, pp. 311–323, 1993.

M. P. Wand, J. S. Marron, and D. Ruppert, "Transformations in density estimation", *Journal of the American Statistical Association*, vol. 86, pp. 343–361, 1991.

Y. G. Yatracos, "Rates of convergence of minimum distance estimators and Kolmogorov's entropy", *Annals of Statistics*, vol. 13, pp. 768–774, 1985.