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

Luc Devroye  
School of Computer Science  
McGill University  
Montreal, Canada H3A 2A7

Gábor Lugosi, and Frederic Udina  
Department of Economics  
Pompeu Fabra University  
Ramon Trias Fargas, 25-27  
08005 Barcelona, Spain

Keywords: Density estimation, kernel estimate, convergence, smoothing factor, minimum distance estimate, asymptotic optimality.

Classification: C13, C14.

The first author's work was supported by NSERC Grant A 3456 and by FCAR Grant 90-ER-0291. The second and third authors' work was supported by DIGES Grant PB96-0300. The third author got lost on the beach while the manuscript was being typed.
ABSTRACT. We continue the development of a method for the selection of a bandwidth or a number of design parameters in density estimation. 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, 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. Further applications include multivariate kernel estimates, transformed kernel estimates, and variable kernel estimates.
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

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

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 $\Theta$ 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 $m < n$, and define $A_{\Theta}$ as the Yatraco 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

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

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

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

over all $\theta \in \Theta$, where $\mu_m$ denotes the empirical measure defined by the subsample $X_{n-m+1}, \ldots, X_n$. 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 $x$, $\mathbb{E}K(x, X) < \infty$.

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

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

This measures the richness of the class of density estimates. The first result upon which many of the other results are built is the following non-asymptotic inequality:
Theorem 1. Let the set $\Theta$ determine a class of regular additive density estimates. Then for all $n, m \leq n/2$, $\Theta$, and $f$,

$$\mathbb{E} \int |f_{n-m, \delta_n} - f| \leq 3 \inf_{\delta \in \Theta} \mathbb{E} \int |f_{n, \delta} - f| \left(1 + \frac{2m}{n-m} + 8 \sqrt{\frac{m}{n}}\right) + \frac{\sqrt{8 \log(4e^s(A_{\Theta}, m^2))}}{m}.$$  

Note that whenever $s(A_{\Theta}, \ell)$ is bounded by a polynomial $n^{k_1 + \ell}$ of $n$ and $\ell$, we have $s(A_{\Theta}, m^2) \leq n^{k_1} m^{2k_2} \leq n^{k_1 + k_2}$, and consequently

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

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

$$\mathbb{E} \int |f_{n-m, \delta_n} - f| \leq 3 \inf_{\delta \in \Theta} \mathbb{E} \int |f_{n, \delta} - f| \left(1 + O \left(1/\sqrt{\log n}\right)\right) + O \left(\frac{\log n}{\sqrt{n}}\right).$$

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.


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 $k$ the Riemann order, which should not be confused with the order of a kernel, which is the smallest positive integer $s$ such that $\int x^s K(x) dx \neq 0$, and is in this sense only defined for univariate kernels.

The standard Akaike-Rosenblatt-Parzen kernel estimate is

$$f_{n,K,h}(x) = f_{n,h}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i),$$
When \( K \) is fixed and \( h \) is chosen by the method described above \((\infty, \theta = h)\), Theorem 1 applies with the following shatter coefficient:

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

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 applies, but now with a slightly larger shatter coefficient:

**Lemma 2.** Consider the class \( \Theta \) 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

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

**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

\[
z_u = \left( \sum_{i=1}^{r} K \left( \frac{y_1 - X_i}{u} \right), \ldots, \sum_{i=1}^{r} K \left( \frac{y_\ell - X_i}{u} \right) \right) \in \mathbb{R}^\ell.
\]

As \( u \uparrow \infty \), each component of \( z_u \) changes every time \((y_j - X_i)/u \) enters or leaves a set \( A_{\ell} \), \( 1 \leq \ell \leq k \) for some \( X_i \), \( 1 \leq i \leq r \), where the \( A_{\ell} \)'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 \( A_{\ell} \), the number of different values a component can take in its history (as \( u \uparrow \infty \)) is clearly bounded by \( 2kr \). As there are \( \ell \) 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 \( \mathcal{W} = \{(w, w') : (w, w') = (z_u, z'_u) \text{ for some } u, v > 0\} \). For fixed \((w, w') \in \mathcal{W}\), let \( U_{(w,w')} \) denote the collection of all \((u, v)\) such that \((z_u, z'_u) = (w, w')\). For \((u, v) \in U_{(w,w')}\), we have

\[
y_i \in A_{u,v} \text{ if and only if } w_i \geq \left(\frac{u}{v}\right)^d w'_i,
\]

where \(w, w'\) have components \(w_i, w'_i\) respectively, \(1 \leq i \leq \ell\). Thus,

\[
\left|\{y_1, \ldots, y_\ell\} \cap A_{u,v} : (u, v) \in U_{(w,w')}\right| \leq \left|\{(I_{w_i \geq c w'_i}, \ldots, I_{w_{\ell} \geq c w'_{\ell}}) : c \geq 0\}\right| \leq \ell + 1.
\]

But then

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

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,

\[
s(A_\Theta, \ell) \leq 4(\ell + 1)(1 + 2k\ell)^2 \leq 8\ell(3k\ell)^2 = 72\ell^3 k^2 r^2.
\]

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

\[
s(A_\Theta, \ell) \leq N^2(\ell + 1)(1 + 2k\ell)^2 \leq 2N^2 \ell(3k\ell)^2 = 18N^2 m^3 k^2 r^2. \quad \square
\]

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 includes kernels of infinite order.

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

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

Note that this is always possible if $K$ is Riemann integrable. The smallest such $k$ will be called the kernel complexity $\kappa_n$. 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

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

is minimal where $A$ is defined as the collection of all sets

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

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 deviation 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 $\kappa_n$. Then for all $n, m \leq n/2$, $d$, and $f$,

$$
\mathbb{E} \int |f_{n-m,K,H} - f| \leq 3 \inf_{h > 0, L \in \mathcal{K}} \mathbb{E} \int |f_{n,L,h} - f| \left( 1 + \frac{2m}{n-m} + 8 \sqrt{\frac{m}{n}} \right) + \sqrt{\frac{8 \log (72 e^8 N^2 \kappa_n^2 n^8)}{m}} + \frac{27}{n}.
$$

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

$$
\mathbb{E} \int |f_{n-m,K,H} - f| \leq 3 \inf_{h > 0, L \in \mathcal{K}} \mathbb{E} \int |f_{n,L,h} - f| \left( 9 + 24/\sqrt{2} \right) + \sqrt{\frac{16 \log (72 e^8 N^2 \kappa_n^2 n^8)}{n}} + \frac{27}{n}.
$$

Proof. Note that

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

Furthermore,

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

Thus, a combination with Theorem 1 then yields, with the appropriate definition of $\Theta$,

$$
\mathbb{E} \int |f_{n-m,K,H} - f| \\
\leq \mathbb{E} \int |f_{n-m,K',H} - f| + \frac{1}{n} \\
\leq 3 \inf_{h > 0, L \in \mathcal{K}} \mathbb{E} \int |f_{n,L,h} - f| \left( 1 + \frac{2m}{n-m} + 8 \sqrt{\frac{m}{n}} \right) + \sqrt{\frac{8 \log (4 e^8 s(\mathcal{A}_\Theta, m^2))}{m}} + \frac{1}{n} \\
\leq 3 \inf_{h > 0, L \in \mathcal{K}} \mathbb{E} \int |f_{n,L,h} - f| \left( 1 + \frac{2m}{n-m} + 8 \sqrt{\frac{m}{n}} \right) + \sqrt{\frac{8 \log (4 e^8 s(\mathcal{A}_\Theta, m^2))}{m}} + \frac{27}{n} \\
\leq 3 \inf_{h > 0, L \in \mathcal{K}} \mathbb{E} \int |f_{n,L,h} - f| \left( 1 + \frac{2m}{n-m} + 8 \sqrt{\frac{m}{n}} \right) + \sqrt{\frac{8 \log (4 e^8 s(\mathcal{A}_\Theta, m^2))}{m}} + \frac{27}{n} \\
\leq 3 \inf_{h > 0, L \in \mathcal{K}} \mathbb{E} \int |f_{n,L,h} - f| \left( 1 + \frac{2m}{n-m} + 8 \sqrt{\frac{m}{n}} \right) + \sqrt{\frac{8 \log (72 e^8 N^2 \kappa_n^2 n^8)}{m}} + \frac{27}{n},
$$

where we used Lemma 2. \(\square\)

The above inequality is useful when $N$ is finite. We will treat the case of infinite kernel classes in the next section. The other quantity of interest is $\kappa_n$. We briefly recall
some bounds from Devroye and Lugosi (1997) and add a useful new bound. Then we conclude by exhibiting an infinite order kernel (with an infinite number of oscillations), yet with finite Riemann order.

Kernels with \( \kappa_n = O(n^b) \) for some finite \( b \) are said to be \textit{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 \( \kappa_n \leq k \). Symmetric unimodal kernels on the real line have \( \kappa_n \leq 8nK(0)\beta + 10 \), where \( \beta \) is the last positive value for which \( \int_\beta^\infty K \leq 1/(4n) \). If \( K(x) \leq aI_{[-b,b]}(x) \) and \( K \) is symmetric, nonnegative, and unimodal (such as the Epanechnikov-Bartlett kernel), then \( \kappa_n \leq 8nab + 10 \). For the normal density, we have \( \kappa_n \leq \frac{8n\sqrt{\log n}}{\sqrt{\pi}} + 10 \). Products of polynomially approximable kernels in \( \mathbb{R}^d \) and multivariate normal densities are also polynomially Riemann approximable. What is of interest in the next section, however, is the following bound.

\textbf{Lemma 3.} 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

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

\textbf{Proof.} Take \( r = 4Dn \) and note that \( \int_{[-r,r]} |K| \leq 2D/r = 1/(2n) \). 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 \( q \geq 2Cr^2n = 32CD^2n^3 \).

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 \( h \),

\[
\liminf_{n \to \infty} n^{s/(2s+1)} \inf_h E \int |f_n, K, h - f| > 0
\]
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 $\alpha$, uniformly over the $N$ kernels,

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

and therefore

$$\sup_{f} \limsup_{n \to \infty} \frac{E \int |f_n, K, H - f|}{\inf_{h} E \int |f_n, K, H - f|} \leq 3 .$$

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

5. Bandwidth selection when the kernel has infinite order.

Watson and Leadbetter (1963) and later, Konakov (1972), Davis (1975, 1977), Ibragimov and Khasminskii (1982), Devroye (1992) and many others documented the usefulness of infinite order kernels $K$ in the Parzen-Rosenblatt estimate when the dimension, $d$, is one. The initial interest was in the sinc kernel $S(x) = \frac{\sin \pi}{\pi}$. The characteristic function of $S$ is $\psi(t) = I_{[-1,1]}(t)$. The flatness of $\psi$ near the origin implies that the best $L_2$ error, $\inf_h E \int (f_n - f)^2$, adapts itself to the density at hand and is thus not a priori limited by the saturation phenomenon seen for finite order kernels. For example, for an order 2 kernel, the best error can decrease at best at a rate $1/n^{4/5}$, whereas with an infinite order kernel, the rate can be $O(1/n)$, which occurs when $f$ has a characteristic function of bounded support. In $L_1$, the results are similar, but the sinc kernel, which is not absolutely integrable, is useless. To obtain an absolutely integrable yet comparable kernel, consider first the trapezoidal characteristic function (Devroye and Győrfi, 1985)

$$\psi(t) = 2(1 - |t|/2) + (1 - |t|) ,$$

a difference between two triangular characteristic functions. As $(1 - |t|)$ is the characteristic function for the de la Vallée-Poussin density

$$L(x) = \frac{1}{2\pi} \left( \frac{\sin(x/2)}{x/2} \right)^2 ,$$

we see that the kernel for $\psi$ is

$$K(x) = 4L(2x) - L(x) = \frac{7 + \cos x - 8 \cos^2 x}{4\pi x^2} .$$

We may apply Lemma 3 here. Indeed, standard calculations and the obvious inequalities $|x - \sin x| \leq x^3/6$ and $|\sin x| \leq |x|$ allow us to verify that $L$ is Lipschitz and that in fact
Thus, although the kernel oscillates infinitely often, it is nevertheless polynomially Riemann approximable. Lemma 3 provides us furthermore with a useful construction for carrying out the computations, as the approximating finite order Riemann kernel is explicitly constructed. Thus, taking \( m = n/2 \) for example, we see that

\[
\mathbb{E} \int |f_{n-m,K,H} - f| \leq 26 \inf_{h>0} \mathbb{E} \int |f_{n,K,H} - f| + D \sqrt{\frac{\log n}{n}}
\]

for a universal constant \( D \). For the vast majority, but not all densities, the last term in the bound is asymptotically negligible with respect to the first term. And, as is well-known, the trapezoidal kernel is optimal (to within a multiplicative factor) within the class of all kernels (see Devroye, 1988, 1992, for an \( L_1 \) treatment of this subject). Densities for which the second term is asymptotically larger include the normal density, where the first term is \( O((\log n)^{1/4}/\sqrt{n}) \), but even in such cases the price paid for complete universality is not more than a factor of \( \sqrt{\log n} \).


In this section, we show by a specific example how one can pick a kernel from an infinite class together with a suitable bandwidth. To make this point, we could consider the univariate class of kernels considered by Hall and Marron in 1988: the kernels are symmetric generalized stable with characteristic functions given by \( \psi(t) = \exp(-|t|^a) \), with \( a \) taking positive even integer values. Note that except for \( a = 2 \), these do not correspond to bona fide densities. This class is interesting and challenging for many reasons. First of all, while \( \int K = 1 \) for all \( a \), we have \( \int |K| \to \infty \) as \( a \to \infty \), because the kernels converge to the sinc kernel (with characteristic function \( L_{-1,1}(t) \)). The number of oscillations varies widely, and the tails of \( K \) become increasingly heavy. Clearly, the bound and methodology of Theorem 2 are not applicable as \( N = \infty \). For utility in \( L_1 \), the class of kernels should have a limit in \( L_1 \), so we modify the Hall-Marron class by considering \( \psi_a(t) = (1 - |t|) \exp(-|t|) \). Then define \( \phi_a(t) = 2\psi_a(t/2) - \psi_a(t) \), and note that for its corresponding kernel \( K_a, \int K_a = 1 \) for all \( a \) and that \( \phi_a \) tends to the trapezoidal characteristic function \( \phi_\infty(t) \equiv 2(1 - |t|/2)_+ - (1 - |t|)_+ \) already encountered earlier.
**Lemma 4.** For \( \phi_a \), with \( a \) even and positive,

\[
\kappa_n \leq 1 + 18432 \xi a^2 n^3
\]

where \( \xi = (6 + \sqrt{2\pi})/\pi < 2.71 \).

**Proof.** First we note that each kernel is in the class covered by Lemma 3. In the remainder of the proof, \( K \) denotes \( K_a \), \( \phi \) denotes \( \phi_a \), and \( \psi \) is \( \psi_a \). As \( K(x) = \frac{1}{2\pi} \int \cos(tx) \phi(t) dt \), we see that

\[
\sup_x |K'(x)| \leq \frac{1}{\pi} \int_0^\infty t|\phi(t)|dt \leq \frac{1}{\pi} \left( 6 + 2 \int_2^\infty te^{-(t/2)^2}dt \right)
\]

\[
\leq \frac{1}{\pi} \left( 6 + 2 \int_2^\infty te^{-(t/2)^2}dt \right) \leq \xi .
\]

Thus, all kernels are Lipschitz with the same Lipschitz constant \( \xi \). Next, let \( L \) be the inverse transform for \( \psi \). Then \( x^2L \) is the inverse transform for \( \psi'' \). Consider \( t > 0 \). Then

\[
\psi''(t) = e^{-t^2}(t^{2a-1}(a^2 + a) - t^{a-2}(a^2 - a) - t^{2a-2}a^2) .
\]

From this, we see that on \([0, 1]\), \( |\psi''| \leq 2a^2e^{-t^2} \) and on \([1, \infty)\), \( |\psi''| \leq 2a^2x^{2a-2}e^{-t^2} \). Using these estimates, we have

\[
\sup_x x^2|L(x)| \leq \int |\psi''(t)|dt \leq \frac{4a^2}{a-1} + 4a^2 \int_1^\infty u^{-1/2}e^{-u}du
\]

\[
\leq \frac{4a^2}{a-1} + \frac{4a^2}{a} \int_1^{\infty} u^{-1/2}e^{-u}du
\]

\[
\leq (4a^2) \left( \frac{1}{a-1} + \frac{1}{a} \right) \leq 12a .
\]

Thus, \( \sup_a x^2|4L(2x) - L(x)| \leq 24a \), where we are making no attempt to optimize the bounds. Thus, Lemma 3 is applicable to \( \phi \) with \( C = \xi \) and \( D = 24a \), so that

\[
\kappa_n \leq 1 + 32 \times 24^2 \xi a^2 n^3 .
\]

The optimization procedure goes as follows. Find Riemann approximations \( K'_a \) for all kernels \( K_a \) in our class \( (K) \) with \( a \) even and \( a \leq 2n^3 \), making sure that \( \int |K_a - K'_a| \leq 1/n \) for all these \( a \). Note that uniformly over \( a \leq 2n^3 \), \( \kappa_n \leq 1 + 73728\xi n^9 \leq \chi n^9 \) with \( \chi = 199804 \geq 1 + 73728\xi \). Let \( \Theta = \{(h, a) : h \geq 0, 2 \leq a \leq 2n^3, a \text{ even}\} \), and pick the best pair \( (H, A) \) from \( \Theta \) as in the previous section, where we picked the best Riemann kernel from a finite class of size \( N \), together with an associated bandwidth. Then the estimate is \( f_{n-m, K_A, H} \). The method is computationally intensive as the Riemann approximants require a lot of work. However, neither \( a \) nor \( h \) are a priori restricted, so that this is a
truly universal method. The performance is bounded uniformly over all densities by the following inequality:

**Theorem 3.** For the method described above, which picks \((H,A)\) from \(\Theta\), we have for all \(n, m \leq n/2\), and \(f\),

\[
E \int |f_{n-m,K_a,H} - f| \leq 3 \inf_{h>0, 2 \leq a \leq 2n^3} E \int |f_{n,K_a,H} - f| \left( 1 + \frac{2m}{n-m} + 8 \sqrt{\frac{m}{n}} \right) + \frac{294 + 256 \log n}{m} + \frac{1053}{n}.
\]

**Proof.** By Theorem 2, if we merely look at the class of kernels \(K_a\) with \(a \leq 2n^3\), we have

\[
E \int |f_{n-m,K_a,H} - f| \leq 3 \inf_{h>0, 2 \leq a \leq 2n^3} E \int |f_{n,K_a,H} - f| \left( 1 + \frac{2m}{n-m} + 8 \sqrt{\frac{m}{n}} \right) + \frac{8 \log \left( 72e^6n^6\chi^2n^{18} \right)}{m} + \frac{27}{n},
\]

where we replaced \(N\) by \(n^3\), and \(\kappa_n\) by its uniform upper bound \(\chi n^9\). To extend the infimum on the left-hand-side of this inequality over all \(a\), we note the following:

\[
\inf_{h>0, 2 \leq a \leq 2n^3} E \int |f_{n,K_a,H} - f| \leq \inf_{h>0, 2 \leq a \leq 2n^3} E \int |f_{n,K_a,H} - f| + \sup_{a \geq 2n^3} \left( E \int |f_{n,K_a,H} - f| - E \int |f_{n,K_{2n^3},H} - f| \right)
\]

\[
\leq \inf_{h>0, 2 \leq a \leq 2n^3} E \int |f_{n,K_a,H} - f| + \sup_{a \geq 2n^3} \int |K_a - K_{2n^3}|
\]

\[
\leq \inf_{h>0, 2 \leq a \leq 2n^3} E \int |f_{n,K_a,H} - f| + 2 \sup_{a \geq 2n^3} \int |K_a - K_\infty|,
\]

where \(K_\infty\) is the limit kernel corresponding to \(\phi_\infty\). By a bound from Lemma 4, \(|K_a(x)| \leq 72a/x^2\), while from the expression for \(K_\infty\), \(|K_\infty(x)| \leq 4/(\pi x^2)\), so

\[
\int_{|x|>r} |K_a - K_\infty| \leq \frac{48a + 8\pi}{r}.
\]

Furthermore,

\[
\int_{|x|\leq r} |K_a - K_\infty| \leq \sqrt{2r} \sqrt{\int |K_a - K_\infty|^2}
\]

(by the Cauchy-Schwartz inequality)
\[ = \sqrt{\frac{r}{\pi}} \int |\phi_a - \phi_\infty|^2 \]
(by Parseval's identity)
\[ \leq \sqrt{\frac{r}{\pi}} 18 \int |\psi_a - \psi_\infty|^2 \]
(by the c inequality, where \( \psi_\infty(t) = (1 - |t|)_+ \)).

For \( 0 \leq t \leq 1 \), we have
\[ |\psi_a(t) - \psi_\infty(t)| = (1 - t)(1 - e^{-t^a}) \leq (1 - t) t^a \leq \left( \frac{a}{a + 1} \right)^a \frac{1}{a + 1} < \frac{1}{a + 1}. \]

For \( 1 \leq t \leq 1 + 1/a \), we have
\[ |\psi_a(t) - \psi_\infty(t)| = (t - 1)e^{-t^a} < \frac{1}{a}. \]

For \( t \geq 1 + 1/a \), the derivative of \( (t - 1)e^{-t^a} \) is nonpositive, so that \( \sup_t |\psi_a - \psi_\infty| \leq 1/a. \) In particular, then,
\[ \int |\psi_a - \psi_\infty|^2 \leq \frac{4}{a^2} + 2 \int_1^\infty (t - 1)^2 e^{-t^a} dt \]
\[ \leq \frac{4}{a^2} + 2e^{-2^a} \int_2^\infty t^2 e^{-t^2} dt \]
\[ \leq \frac{4}{a^2} + e^{-2^a} < \frac{5}{a^2}. \]

Combining all this shows that
\[ \int |K_a - K_\infty| \leq \frac{48a + 8/\pi}{r} + \sqrt{\frac{90r}{\pi a^2}} < \frac{50a}{r} + \sqrt{\frac{30r}{a^2}} \]
\[ = \frac{10}{a^{1/3}} + \frac{\sqrt{150}}{a^{1/3}} \quad \text{(take } r = 5a^{4/3}) \]
\[ < \frac{23}{a^{1/3}} \leq \frac{19}{n} \]
as \( a \geq 2n^3 \). Thus,
\[ \inf_{h>0, 0 \leq a < 2n^3} \mathbb{E} \int |f_n, K_s H - f| \leq \inf_{h>0, 0 \leq a < 2n^3} \mathbb{E} \int |f_n, K_s H - f| + \frac{38}{n}. \]

Finally,
\[ \mathbb{E} \int |f_n - m, K_s H - f| \leq \frac{3}{n} \inf_{h>0, 0 \leq a < 2n^3} \mathbb{E} \int |f_n, K_s H - f| \left( 1 + \frac{2m}{n - m} + 8\sqrt{\frac{m}{n}} \right) \]
\[ + \sqrt{\frac{81}{m}} \left( 72 e^{8n^6 \chi^2 n^{18} n^8} \right) \frac{27}{n} + \frac{3 \times 38 \times 9}{n}. \]
Theorem 3 now follows as \(8 \log(72e^8 \chi^2) < 294\). □

The constants in Theorem 3 are of course not best possible, but we only wanted to work out one example in all its details, to show that the algorithm has indeed the required asymptotic performance. Undoubtedly, in practice, one either uses one infinite order kernel or optimizes over just a few low order kernels. As shown above, nothing keeps one from optimizing over an infinite class of kernels. In fact, it is of independent interest to design such infinite classes with computational efficiency in mind.

7. 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 5. Assume that for each \( j, K_j = I_{A_j} \), where \( A_j = [a_j, a_j + 1] \) is an interval of unit length. Then for \( \ell n \geq 2d \) we have
\[
s(A_{\Theta}, \ell) \leq (\ell + 1) \left( \frac{\ell n e}{2d} \right)^{4d}.
\]

Lemma 5 shows that the shatter coefficient is polynomial in \( n \) and \( \ell \). 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 5 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 5. 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
\[
z_\theta = (z_\theta^{(1)}, \ldots, z_\theta^{(\ell)}) = \left( \sum_{i=1}^{n-m} \prod_{j=1}^{d} K_j \left( \frac{y_t^{(j)} - X_i^{(j)}}{h_j} \right), \ldots, \sum_{i=1}^{n-m} \prod_{j=1}^{d} K_j \left( \frac{y_t^{(j)} - X_i^{(j)}}{h_j} \right) \right).
\]
Observe that for each \( t \leq \ell \) and \( i \leq n - m \)
\[
\prod_{j=1}^{d} K_j \left( \frac{y_t^{(j)} - X_i^{(j)}}{h_j} \right) = 1 \quad \text{if and only if} \quad y_t - X_i \in R_\theta,
\]
where \( R_\theta \) 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 \( z_\theta \) can take as \( \theta \) varies through \( \Theta \) is at most \( s(B, \ell(n - m)) \), where \( B \) 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 \( s(B, \ell(n - m)) \leq \left( \frac{\ell(n - m)e}{2d} \right)^{2d} \).

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 \((\theta_1, \theta_2)\) such that \((z_{\theta_1}, z_{\theta_2}) = (w, w')\). For \((\theta_1, \theta_2) \in U_{(w, w')}\), we have
\[
y_t \in A_{\theta_1, \theta_2} \quad \text{if and only if} \quad z_\theta^{(t)} \prod_{j=1}^{d} \frac{1}{h_{j,1}} \geq z_{\theta_2}^{(t)} \prod_{j=1}^{d} \frac{1}{h_{j,2}},
\]
17
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
\[
\left| \{ y_1, \ldots, y_t \} \cap A_{w,v} : (u,v) \in U_{(w,w')} \right| \leq \left| \{ (I_{w_1 \leq c w_1}, \ldots, I_{w_2 \geq c w_2}) : c \geq 0 \} \right| \leq \ell + 1,
\]
where \( w_1, \ldots, w_t \) and \( w'_1, \ldots, w'_t \) denote the components of the vectors \( w \) and \( w' \), respectively. But then
\[
\left| \{ y_1, \ldots, y_t \} \cap A_{w,v} : (u,v) > 0 \right| \leq (\ell + 1) \left| U_{(w,w')} \right| \leq (\ell + 1) \left( \frac{\ell (n - m) e}{2d} \right)^{4d}.
\]


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 \( \Sigma \) is a positive definite symmetric \( d \times d \) matrix, and
\[
K_{\theta}(x) = v_\theta I_{(x^T \Sigma^{-1} x \leq 1)}.
\]
Here \( v_\theta \) is a normalizing factor such that \( \int K_{\theta} = 1 \), and \( x^T \) denotes the transpose of the vector \( x \). In this case, for \( \ell(n - m) \geq d^2 + d + 2 \), we have
\[
s(A_\theta, \ell) \leq (\ell + 1) \left( \frac{\ell (n - m) e}{d^2/2 + d/2 + 1} \right)^{d^2 + d + 2}.
\]

The proof is exactly the same as for the case of product kernels with the only difference that the shatter coefficients of the class \( E \) of ellipsoids (i.e., class of sets of the form \( E_\theta = \{ x : x^T \Sigma^{-1} x \leq 1 \} \)) is bounded by
\[
s(E, \ell(n - m)) \leq \left( \frac{\ell (n - m) e}{d^2/2 + d/2 + 1} \right)^{d^2 + d/2 + 1},
\]
whenever \( \ell(n - m) \geq d^2 + d + 2 \) (since the \( \text{VC} \) dimension of \( E \) is bounded by \( d^2/2 + d/2 + 1 \), see, e.g., Devroye, Gyórfi, and Lugosi (1996, p.221). Although it is computationally challenging to optimize all \( \binom{d}{d} \) 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.

Variable kernel estimates use different smoothing factors in different regions of $\mathbb{R}^d$. To obtain good estimates, the smoothing factors and the regions should be selected based on the data. More precisely, if $P = \{B_1, B_2, \ldots, B_r\}$ is a partition of $\mathbb{R}^d$ into $r$ sets, then the corresponding variable kernel estimate is defined as

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

where $h = (h_1, \ldots, h_r) \in (\mathbb{R}^+)^r$ is a vector of positive smoothing factors, and $\theta = (P, h)$. Clearly, $\int K = 1$ implies that $\int f_{n, \theta} = 1$. Let $\mathcal{P}$ be a family of partitions of $\mathbb{R}^d$ such that each $P \in \mathcal{P}$ has at most $r$ cells. Let $\Theta = \{(P, h) : P \in \mathcal{P}, h \in (\mathbb{R}^+)^r\}$ denote the set of all possible pairs of partitions and vectors of smoothing factors.

With a slight abuse of notation, denote by $s(\mathcal{P}, n)$ the $n$-th shatter coefficient of the class of sets which are cells of any partition in $\mathcal{P}$. Again, for simplicity, we assume that $K$ is the window kernel $K(x) = cI_{\{\|x\| \leq 1\}}$, where $c$ is an appropriate normalizing factor.

**Lemma 6 (Cover (1965)).** Let $A$ 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

$$s(A, n) \leq 2 \sum_{i=0}^{d} \binom{n-1}{i} \leq 2(n-1)^d + 2 \leq 2n^d.$$
**Lemma 7.** If $\mathcal{A}_\Theta$ is the Yatracos class defined by $\Theta = \{ (P, h) : P \in \mathcal{P}, h \in (\mathbb{R}^+)^r \}$, then

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

In particular, if $d = 1$, and $\mathcal{P}$ is the class containing all partitions of the real line into at most $r$ intervals, then $s(\mathcal{P}, n)$ is just the $n$-th shatter coefficient of the class of all intervals, which is well-known to be equal to $n(n+1)/2 + 1$ (see, e.g., Devroye, Györfi, and Lugosi (1995, p.220), and therefore

$$s(\mathcal{A}_\Theta, \ell) \leq 2\ell^{2r} (\ell^2(n-m)^2)^{2r} (2\ell(n-m))^{2r} \leq 2^{1+2r} \ell^{2r} n^{6r}.$$ 

For fixed $r$, this bound is polynomial in $n$ and $\ell$, and the remarks of the previous sections apply. If $r$ is allowed to tend to infinity, then the situation is different, as the additive term in the upper bounds of the $L_1$ error is proportional to $\sqrt{r}$. At some point, when $r$ increases too fast, the additive term becomes too large, and we lose the asymptotic optimality modulo 3. It should be of interest to investigate the possibility or impossibility of finding data-based choices for the variable kernel estimate in which each $X_i$ has its own tunable $H_i$, as in the original estimate of Breiman, Meisel and Purcell (1977). It certainly seems as though one cannot obtain universal inequalities in which the additive term is asymptotically smaller than the multiplicative term. Hence, the need to restrict the complexity of the class of tunable parameters as we did in this section.

**Proof of Lemma 7.** For each $\theta \in \Theta$, consider the $\ell \times r$ matrix $z_\theta$ such that the element in its $t$-th row and $j$-th column is

$$z^{(t,j)}_\theta = \sum_{i=1}^{n-m} I_S \left( \frac{y_t - X_i}{h_j} \right) I_{B_j}(X_i), \quad t \leq \ell, j \leq r,$$

where $S = \{ x : \|x\| \leq 1 \}$ is the closed unit ball centered at the origin. Clearly,

$$I_S \left( \frac{y_t - X_i}{h_j} \right) I_{B_j}(X_i) = 1 \quad \text{if and only if} \quad (X_i, y_t - X_i) \in B_j \times S(0, h_j),$$

where $S(0, h) = \{ x : \|x\| \leq h \}$. Since there are $\ell(n-m)$ different pairs $(X_i, y_t - X_i)$, the number of different values the $j$-th column $(z^{(1,j)}_\theta, \ldots, z^{(\ell,j)}_\theta)$ of the matrix $z_\theta$ can take as $\theta \in \Theta$ is at most the shatter coefficient $s(C, \ell(n-m))$ of the class of sets $C$ of the form $B \times S(0, h)$, where $h$ ranges through the set of positive numbers, and $B$ is any set which can be a cell of a partition in $\mathcal{P}$. This shatter coefficient is clearly bounded by the product of the $\ell(n-m)$-th shatter coefficient of the class of sets $\{ S(0, h) : h > 0 \}$ and the shatter coefficient $s(\mathcal{P}, \ell(n-m))$ of the class of sets which are cells of any partition in $\mathcal{P}$. Since the $\ell(n-m)$-th shatter coefficient of $\{ S(0, h) : h > 0 \}$ is $\ell(n-m) + 1$, the $j$-th column of $z_\theta$ can take at most $(\ell(n-m) + 1)s(\mathcal{P}, \ell(n-m))$ values. But since the matrix $z_\theta$ has
columns, it can take at most

\[((\ell(n - m) + 1)s(P, \ell(n - m)))^r\]

values. Thus,

\[|\{(z_{\theta_1}, z_{\theta_2}) : \theta_1, \theta_2 \in \Theta\}| \le ((\ell(n - m) + 1)s(P, \ell(n - m)))^{2r}.\]

Let \(W = \{(w, w') : (w, w') = (z_{\theta_1}, z_{\theta_2})\text{ for some } \theta_1, \theta_2 \in \Theta\}\). For fixed \((w, w') \in W\), let \(U_{(w, w')}\) denote the collection of all \((\theta_1, \theta_2)\) such that \((z_{\theta_1}, z_{\theta_2}) = (w, w')\). For \((\theta_1, \theta_2) \in U_{(w, w')}\), and \(t \le \ell\) we have

\[y_t \in A_{\theta_1, \theta_2}\text{ if and only if } \sum_{j=1}^r \frac{1}{h_{j,1}}z_{\theta_1}^{(t,j)} \ge \sum_{j=1}^r \frac{1}{h_{j,2}}z_{\theta_2}^{(t,j)},\]

where \(\theta_1 = (h_{1,1}, \ldots, h_{r,1})\) and \(\theta_2 = (h_{1,2}, \ldots, h_{r,2})\). Within the set \(U_{(w, w')}\), \(z_{\theta_1}^{(t,j)}\) and \(z_{\theta_2}^{(t,j)}\) are fixed for all \(t\) and \(j\), and therefore the number of different values of the binary vector

\[\left(IA_{\theta_1, \theta_2}(y_1), \ldots, IA_{\theta_1, \theta_2}(y_{\ell})\right)\]

is at most the number of different values the vector

\[\left(I_{\sum_{j=1}^{r-j}(\frac{1}{h_{j,1}}z_{\theta_1}^{(1,j)} - \frac{1}{h_{j,2}}z_{\theta_2}^{(1,j)}) \ge 0}, \ldots, I_{\sum_{j=1}^{r-j}(\frac{1}{h_{j,1}}z_{\theta_1}^{r-j} - \frac{1}{h_{j,2}}z_{\theta_2}^{r-j}) \ge 0}\right)\]

takes as \(h_{1,1}, \ldots, h_{r,1}, h_{1,2}, \ldots, h_{r,2}\) all vary through \(\mathbb{R}^+\). But this is not more than the maximal number of different ways of dichotomizing \(\ell\) points in \(\mathbb{R}^{2r}\) by hyperplanes, which, by Lemma 6, is at most \(2\ell^{2r}\). The proof of Lemma 7 is finished. \(\square\)
10. 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. In our second example, the transformation is restricted to be piecewise linear consisting of a fixed number of pieces, but otherwise it is 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 $T : \mathbb{R} \rightarrow \mathbb{R}$ 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 possibilities, 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 transformation (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

\[
f_{n,\theta}(x) = \frac{1}{nh} \sum_{i=1}^{n} I_{[\{ T_\theta(x_i) - T_\theta(X_t) \leq h/2 \} \cap \mathbb{I}^a]} x^{a-1},
\]

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 \( s(A_\Theta, \ell) \).

**Lemma 8.** Let \( A_\Theta \) denote the Yatracos class corresponding to the family of kernel estimates on \( \mathbb{IR}^+ \) based on all Box-Cox transformations \( T_\theta \), \( a \in [0, 1] \) and all smoothing factors \( h > 0 \). If \( \ell \geq 2 \) and \( n - m \geq 2 \), then

\[
s(A_\Theta, \ell) \leq \frac{9}{4} \ell^6 (n - m)^a.
\]

**Proof.** In the proof we use a simple lemma which is an easy modification of Lemma 25.2 of Devroye, Györfi, and Lugosi (1995):

**Lemma 9.** If \( b_1, \ldots, b_k, c_1, \ldots, c_k \in \mathbb{IR} \), 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.

**Proof of Lemma 8.** Consider the \( \ell \times (n - m) \) matrix \( z_{a,h} \) with entries

\[
z_{a,h}^{t,i} = I_{[\{ X_t - y_i \} \leq ah]} \quad t = 1, \ldots, \ell, \quad i = 1, \ldots, n - m.
\]

First we bound the number of different possible 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

\[
u^a - z^a - ah = 0 \quad \text{where} \quad u = \max(X_t, y_i) \quad \text{and} \quad v = \min(X_t, y_i).
\]

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

\[
e^{alog u} - e^{alog v} - e^{alog w} + e^{alog z} = 0.
\]

According to Lemma 9, 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 \( \theta' = (a', h') \),

\[
y_i \in A_{\theta, \theta'} \quad \text{if and only if} \quad \frac{y_i^{a-1} \sum_{i=1}^{n-m} w^{(t,i)}}{h} \geq \frac{y_i^{a'-1} \sum_{i=1}^{n-m} w'^{(t,i)}}{h'},
\]
or equivalently, if and only if

\[
(a - 1) \log y_i - \log h + \log \left( \sum_{i=1}^{n-m} w^{(t,i)} \right) \geq \left( a' - 1 \right) \log y_i - \log h' + \log \left( \sum_{i=1}^{n-m} w'^{(t,i)} \right).
\]

Therefore, denoting \( W_i = \sum_{i=1}^{n-m} w^{(t,i)} \) and \( W'_i = \sum_{i=1}^{n-m} w'^{(t,i)} \), 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

\[
(I(a-a') \log y_i - \log(h/h') + \log(W_i/W'_i) \geq 0) \quad \cdots \quad (I(a-a') \log y_i - \log(h/h') + \log(W_i/W'_i) \geq 0)
\]
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 \( \ell \) points by 2-dimensional hyperplanes, which, by Lemma
6, is at most \( \ell^2 \) (since \( \ell \geq 2 \)). Collecting bounds, the proof is finished. \( \Box \)

Having Lemma 8, Theorem 1 yields the following bound:
Theorem 4. 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

\[
f_{n,a,h}(x) = \frac{1}{nh} \sum_{i=1}^{n} I_i \{ |T_a(x_i) - T_a(x)| < h/2 \} x^{a-1}.
\]

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

\[
E \int |f_n - f| \leq 3 \sup_{a \in [0,1], h > 0} E \int |f_{n,a,h} - f| \left( 1 + \frac{2m}{n - m} + 8 \sqrt{\frac{m}{n}} \right) + \sqrt{ \frac{8 \log (9e^8m^{12}(n-m)^4)}{m} }.
\]

For example, if \( n \) is even and we take \( m = n/2 \),

\[
E \int |f_n - f| \leq 26 \sup_{a \in [0,1], h > 0} E \int |f_{n,a,h} - f| + 16 \sqrt{ \frac{\log n}{n} }.
\]

Piecewise linear transformations. Here we assume that each \( T \) in the collection is piecewise linear with at most \( r \) pieces. More precisely, \( T \) has the form

\[
T(x) = \sum_{j=1}^{r} (a_j x + b_j) I_{B_j}(x),
\]

where \( B_1, \ldots, B_r \) is a partition of \( \mathbb{R} \) into \( r \) intervals defined by the endpoints \( c_1 \leq \cdots \leq c_{r-1} \in \mathbb{R} \), and \( a_1, \ldots, a_r > 0, b_1, \ldots, b_r \in \mathbb{R} \) are real constants. Assume for simplicity that \( K = I_{[-1/2, 1/2]} \) is the window kernel.

Lemma 10. If \( \Theta \) denotes the set of all possible \( 3r-1 \)-dimensional vectors \( (a_1, \ldots, a_r, b_1, \ldots, b_r, c_1, \ldots, c_{r-1}) \), and \( \ell(n-m) \geq 2 \), then

\[
s(\mathcal{A}_\Theta, \ell) \leq 2\ell^{2r} \left[ (\ell + n - m - 1) ((n-m)\ell)^{4r^2 - 2} \right] \leq 2\ell^{8r^2 + 4r} n^{8r^2 + 2r}.
\]
Proof. For each \( \theta \in \Theta \), we define an \( \ell \times r \) matrix \( z_\theta \) with entries

\[
z_\theta^{(t,j)} = I_{B_j}(y_t) \sum_{i=1}^{n-m} I_{[-1/2,1/2]} \left( a_j y_t + b_j - \sum_{j'=1}^{r} (a_{j'} X_i + b_{j'}) I_{B_j}(X_i) \right), \quad t \leq \ell, j \leq r.
\]

First we bound the number of different values the matrix \( z_\theta \) can take as \( \theta \) ranges through \( \Theta \). First consider the \( r(\ell + (n - m)) \)-vector of the indicators \( I_{B_j}(y_t), I_{B_j}(X_i), t \leq \ell, i \leq n - m, j \leq r \). Clearly, the number of different values this vector can take is not more than the number of different ways \( \ell + (n - m) \) points can be partitioned into \( r \) intervals. This is at most \((\ell + n - m - 1)\). Consider now a set of partitions into \( r \) intervals such that the indicators \( I_{B_j}(y_t) \) and \( I_{B_j}(X_i) \) are fixed \( t \leq \ell, i \leq n - m, j \leq r \). By the above argument there are at most \((\ell + n - m - 1)\) such sets. Over such a set, the matrix \( z_\theta \) cannot take more distinct values than the \((n - m) \times \ell\) matrix with entries

\[
I_{[-1/2,1/2]}(a_j y_t + b_j - a_{j(i)} X_i - b_{j(i)}) \quad i \leq n - m, t \leq \ell,
\]

where \( j_t \) is the index of the set \( B_j \) containing \( y_t \), and \( j(i) \) is the index of the set \( B_j \) containing \( X_i \). But since the number of different values of the pair \((j_t, j(i))\) is at most \( r^2 \), the number of values the above matrix can take does not exceed the number of different ways \((n - m)\ell\) two-dimensional points (formed by the pairs \((X_i, y_t)\)) can be dichotomized by \( 2r^2 \) hyperplanes, which, by Lemma 6 (since \( \ell(n - m) \geq 2 \)), is at most \((n - m)\ell/4r^2\). Therefore, the number of different values of the matrix \( z_\theta \) is at most \((\ell + n - m - 1)/4r^2\).

From this point on we proceed exactly as in the previous sections. Thus,

\[
\left| \{(z_{\theta_1}, z_{\theta_2}) : \theta_1, \theta_2 \in \Theta \} \right| \leq \left( \frac{\ell + n - m - 1}{r - 1} \right)^2 \left(\frac{(n - m)\ell}{4r^2}\right)^2.
\]

Let \( W = \{(w, w') : (w, w') = (z_{\theta_1}, z_{\theta_2}) \) for some \( \theta_1, \theta_2 \in \Theta \}. \) For fixed \((w, w') \in W\), let \( U_{(w,w')}\) denote the collection of all \((\theta_1, \theta_2)\) such that \((z_{\theta_1}, z_{\theta_2}) = (w, w')\). For \((\theta_1, \theta_2) \in U_{(w,w')}\), we have

\[
y_t \in A_{\theta_1, \theta_2} \quad \text{if and only if} \quad \sum_{j=1}^{r} a_{j,1} z_{\theta_1}^{(t,j)} + \sum_{j=1}^{r} a_{j,2} z_{\theta_2}^{(t,j)},
\]

where \( \theta_i = (a_{i,1}, \ldots, a_{i,r}, b_{1,i}, \ldots, b_{r,i}, c_{1,i}, \ldots, c_{r-1,i}) \) for \( i = 1, 2 \). Within the set \( U_{(w,w')}\), \( z_{\theta_1}^{(t,j)} \) and \( z_{\theta_2}^{(t,j)} \) are fixed for all \( t \) and \( j \), and therefore the number of different values of the binary vector

\[
(I_{A_{\theta_1, \theta_2}}(y_1), \ldots, I_{A_{\theta_1, \theta_2}}(y_{\ell}))
\]

is at most the number of different values of vector

\[
\left( I_{\sum_{j=1}^{r} \sum_{j=1}^{r} (a_{j,1} z_{\theta_1}^{(t,j)} - a_{j,2} z_{\theta_2}^{(t,j)}) \geq 0), \ldots, I_{\sum_{j=1}^{r} \sum_{j=1}^{r} (a_{j,1} z_{\theta_1}^{(t,j)} - a_{j,2} z_{\theta_2}^{(t,j)}) \geq 0} \right)
\]

26
takes as \( a_{1,1}, \ldots, a_{r,1}, a_{1,2}, \ldots, a_{r,2} \) all vary through \( \mathbb{R}^+ \). But this is not more than the maximal number of different ways of dichotomizing \( \ell \) points in \( \mathbb{R}^{2r} \) by hyperplanes, which, according to Lemma 6, is at most \( 2\ell^{2r} \). The proof of is finished. \( \square \)

11. Uniform bounds over classes of densities.

Assume that we have a class of densities \( \mathcal{F} \). We define the minimax risk \( R_n(\mathcal{F}) \) by

\[
R_n(\mathcal{F}) = \inf_{f_n} \sup_{f \in \mathcal{F}} \mathbb{E} \int |f_n - f|
\]

where \( f_n \) is any density estimate. We say that a given density estimate is minimax optimal if

\[
\sup_{f \in \mathcal{F}} \mathbb{E} \int |f_n - f| \leq C R_n(\mathcal{F})
\]

for some finite constant \( C \) not depending upon \( n \). Surveys of recent results may be found in Devroye (1987), devroye and Györfi (1985), Yang (1996), with relevant \( L_1 \) material also available in the work of Birgé (1985) and Bretagnolle and Huber (1977). Our results are useful in the following sense. Assume that for a parametrized class \( f_{n,\theta} \), with \( \theta \in \Theta \),

\[
\sup_{f \in \mathcal{F}} \inf_{\theta \in \Theta} \mathbb{E} \int |f_{n,\theta} - f| \leq C' R_n(\mathcal{F})
\]

for some constant \( C' \) and all \( n \). The choice of \( \theta \) depends upon the density \( f \) and is thus not available. However, for some estimates in this paper, we can find a data-based selection of the parameters (denoted by \( \Theta_n \)) such that for another constant \( C'' \),

\[
\sup_{f \in \mathcal{F}} \mathbb{E} \int |f_{n,\theta_n} - f| \leq C'' R_n(\mathcal{F})
\]

This, of course, is entirely due to the non-asymptotic nature of our bounds, and the fact that the bounds are uniform over all densities. One needs only calculate the shatter coefficient for the class of estimates at hand, and verify that the additive term in the bound is asymptotically negligible. While we will work out the details for some examples elsewhere, it is of interest to note that for the kernel estimate, with \( \Theta \) the space of all \( h > 0 \) and with fixed superkernel \( K \), the \( C' \)-inequality above holds when \( \mathcal{F} \) is a typical smoothness class, such as any class of densities on \([0,1]\) in which \( f \) has \( s - 1 \) absolutely continuous derivatives, and \( f^{(s)} \) is Lipschitz with Lipschitz constant \( (C) \). The same \( \Theta \) can be used for all \( s \) simultaneously! Therefore, the data-based method described here yields a kernel estimate that adapts itself to any order \( s \) of the class, and indeed to every Lipschitz constant \( C \) in the definition of \( \mathcal{F} \). Details and explicit bounds will be worked out elsewhere, both for smoothness classes and for classes that are defined in terms of their tails.

Lemma 11.

\[ \int |f_{n-m,\theta_n} - f| \leq 3 \inf_{\theta \in \Theta} \int |f_{n-m,\theta} - f| + 4 \sup_{A \in A_\Theta} \left| \int_A f - \mu_m(A) \right|. \]

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

\[ \int |\tilde{f} - f| \leq \int |f_{n-m,\theta} - f| + \epsilon \]

for all \( \theta \in \Theta \). Then

\[ \int |f_{n-m,\theta_n} - f| \leq \int |\tilde{f} - f| + \int |f_{n-m,\theta_n} - \tilde{f}| \]

\[ = \int |\tilde{f} - f| + 2 \sup_{A \in A_\Theta} \left| \int_A f_{n-m,\theta_n} - \int_A \tilde{f} \right| \quad \text{(by Scheffé's theorem)}, \]

\[ \leq \int |\tilde{f} - f| + 2 \sup_{A \in A_\Theta} \left| \int_A f_{n-m,\theta_n} - \mu_m(A) \right| + 2 \sup_{A \in A_\Theta} \left| \mu_m(A) - \int_A \tilde{f} \right| \quad \text{(by the triangle inequality)} \]

\[ \leq \int |\tilde{f} - f| + 4 \sup_{A \in A_\Theta} \left| \mu_m(A) - \int_A \tilde{f} \right| \quad \text{(by the definition of \( \theta_n \))} \]

\[ \leq \int |\tilde{f} - f| + 4 \sup_{A \in A_\Theta} \left| \int f - \int \tilde{f} \right| + 4 \sup_{A \in A_\Theta} \left| \mu_m(A) - \int_A f \right| \quad \text{(by the triangle inequality)} \]

\[ \leq 3 \inf_{\theta \in \Theta} \int |f_{n-m,\theta} - f| + 3 \epsilon + 4 \sup_{A \in A_\Theta} \left| \int_A f - \mu_m(A) \right|. \]

Since \( \epsilon \) is arbitrary, Lemma 11 is proved. \( \square \)

The following simple lemmas are used in the proof:

Lemma 12. Let \( X \) and \( Y \) be independent random variables, and let \( \mathbb{E} Y = 0 \). Then \( \mathbb{E}|X + Y| \geq \mathbb{E}|X| \).
**Lemma 13** (Devroye and Györfi, 1985, page 137). Let $Y_1, \ldots, Y_n$ be i.i.d. zero mean random variables. Then

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

**Lemma 14**. Let $\Theta$ 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

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

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

$$
\inf_{\theta \in \Theta} E J_{n-m,\theta} \leq \inf_{\theta \in \Theta} E J_{n,\theta} \times \sup_{\theta \in \Theta} \left( \frac{E J_{n-m,\theta}}{E J_{n,\theta}} \right)
$$

The supremum is rewritten as follows:

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

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

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

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 $Y_1$ 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

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

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

$$
\frac{E|f_{n-m,\theta} - f_{n,\theta}|}{E|f_{n,\theta} - E f_{n,\theta}|} \leq \frac{m}{n-m} \frac{E|S_{n-m}| + E|S_m|}{E|S_n|} \leq \frac{m}{n-m} + \frac{E|S_m|}{E|S_n|} \quad \text{(because $E|S_n| \geq E|S_{n-m}|$)}
$$

29
\[
\frac{m}{n - m} + \frac{\mathbb{E}|S_m|}{\sqrt{\frac{n/m}{\mathbb{E}|S_m|}}} \quad \text{(by Lemmas 12 and 13)}
\]
\[
\leq \frac{m}{n - m} + 4\sqrt{\frac{m}{n}} \quad \text{(if } 2m \leq n) .
\]

This implies that for any fixed \( \theta \),
\[
\mathbb{E} \int |f_{n-m, \theta} - f_{n, \theta}| \, d\mathbf{x} \leq \left( \frac{m}{n - m} + 4\sqrt{\frac{m}{n}} \right) \mathbb{E} \int |f_{n, \theta} - \mathbb{E} f_{n, \theta}| \, d\mathbf{x} .
\]

The lemma now follows without work. \( \square \)

We now note the following: a variant of the Vapnik-Chervonenkis inequality (Vapnik and Chervonenkis (1971); see Devroye (1982)) states that for \( \epsilon > 0 \),
\[
\mathbb{P} \left\{ \sup_{A \in \mathcal{A}_0} \left| \mu_m(A) - \int_A f \right| > \epsilon \Big| X_1, \ldots, X_{n-m} \right\} \leq 4\epsilon^8 s(A_0, m^2) e^{-2m^2}.
\]

This implies by standard bounding that
\[
\mathbb{E} \left\{ \sup_{A \in \mathcal{A}_0} \left| \mu_m(A) - \int_A f \right| \Big| X_1, \ldots, X_{n-m} \right\} \leq \sqrt{\frac{\log \left( 4\epsilon^8 s(A_0, m^2) \right)}{2m}}
\]
(see Devroye, Györfi, and Lugosi (1995, page 208)). Theorem 1 now follows from this estimate, Lemma 11 and Lemma 14.

13. References


