The Annals of Statistics
1996, Vol. 24, No. 2, 508–539

DENSITY ESTIMATION BY WAVELET THRESHOLDING[1]

BY DAVID L. DONOHO, IAIN M. JOHNSTONE,
GÉRARD KERKYACHARIAN AND
DOMINIQUE PICARD

Stanford University, Université de Picardie and Université de Paris VII

Density estimation is a commonly used test case for nonparametric estimation methods. We explore the asymptotic properties of estimators based on thresholding of empirical wavelet coefficients. Minimax rates of convergence are studied over a large range of Besov function classes $B_{\sigma pq}$ and for a range of global L'_p error measures, $1 \le p' < \infty$. A single wavelet threshold estimator is asymptotically minimax within logarithmic terms simultaneously over a range of spaces and error measures. In particular, when $p' > p$, some form of nonlinearity is essential, since the minimax linear estimators are suboptimal by polynomial powers of n. A second approach, using an approximation of a Gaussian white-noise model in a Mallows metric, is used to attain exactly optimal rates of convergence for quadratic error ($p' = 2$).

1. Introduction. The recent appearance of explicit orthonormal bases based on multiresolution analyses has exciting implications for nonparametric function estimation. Unlike the traditional Fourier bases, wavelet bases offer a degree of localization in space as well as frequency. This enables development of simple function estimates that respond effectively to discontinuities and spatially varying degrees of oscillations in a signal, even when the observations are contaminated by noise.

This paper applies these heuristics to probability density estimation: estimate a probability density function $f(x)$ on the basis of $X_1, \ldots, X_n$, independent and identically distributed observations drawn from f. Because of its simplicity, this important practical problem has also served as one of the basic test situations for the theory of nonparametric estimation. An overview of traditional methods and a part of the vast literature on theory and application of density estimation is given by Devroye and Györfi (1985), Silverman (1986) and Scott (1992). The first use of wavelet bases for density estimation appears in papers by Doukhan and Leon (1990), Kerkyacharian and Picard (1992) and Walter (1992).

Let us suppose that the (inhomogeneous) wavelet basis is derived from $\{\phi_{j_1,k} = 2^{j_1/2}\phi(2^{j_1}x - k),\ k \in \mathbb{Z}\}$ and $\{\psi_{jk} = 2^{j/2}\psi(2^j x - k),\ k \in \mathbb{Z},\ j \ge j_1\}$, where $\phi(x)$ and $\psi(x)$ are the scaling function and mother wavelet, respectively.

Received April 1993; revised May 1995.

[1]This work was supported in part by NSF Grant DMS-92-09130 and NIH PHS Grant GM21215-12.

AMS 1991 *subject classifications*. 62G07, 62G20.

Key words and phrases. Minimax estimation, adaptive estimation, density estimation, spatial adaptation, wavelet orthonormal bases, Besov spaces.

The probability density f has formal expansion

$$f(x) \sim \sum_k \alpha_{j_1 k} \phi_{j_1 k}(x) + \sum_{j \geq j_1} \sum_k \beta_{jk} \psi_{jk}(x). \tag{1}$$

Since wavelet estimators are a form of orthogonal series estimate, one begins by forming empirical wavelet coefficients

$$\hat{\alpha}_{j_1, k} = n^{-1} \sum_{i=1}^{n} \phi_{j_1 k}(X_i), \qquad \hat{\beta}_{jk} = n^{-1} \sum_{i=1}^{n} \psi_{jk}(X_i). \tag{2}$$

The key advantages of wavelet estimators follow from the effects of even very simple nonlinearities involving coordinatewise thresholding:

$$\delta_s(x, \lambda) = \operatorname{sgn} x(x - \lambda)_+, \qquad \delta_h(x, \lambda) = xI\{|x| > \lambda\},$$

where the subscripts refer to "soft" and "hard" thresholding, respectively. The estimators we consider in this paper are obtained by thresholding empirical coefficients:

$$\tilde{\beta}_{jk} = \delta(\hat{\beta}_{jk}, \lambda_j), \qquad \delta = \delta_s, \delta_h, \tag{3}$$

and using $\tilde{\beta}_{jk}$ along with $\hat{\alpha}_{j_1 k}$ in (1) to form the estimate $\hat{f}_n$. Here we use either soft or hard thresholding as dictated by technical convenience—from simulation experience in other contexts, one expects that soft thresholding will better suppress noise artifacts, while hard thresholding will better preserve the visual appearance of peaks and jumps.

We look at global error measures for estimating the whole density, evaluating the mean $L_{p'}$ error

$$R_n(\hat{f}, f) = E\|\hat{f}_n - f\|_{p'}^{p'} = E \int |f_n(x) - f(x)|^{p'} \, dx \quad \text{for } 1 \leq p' < \infty$$

and

$$R_n(\hat{f}, f) = E\|f_n(x) - f(x)\|_\infty \quad \text{for } p' = \infty.$$

For the most part, we consider $1 \leq p' \leq \infty$, which includes the important special cases $p' = 1$ and 2, which are of interest, respectively, for their properties of invariance and mathematical simplicity. We look at the worst case performance over a variety of functional spaces:

$$R_n(\hat{f}; \mathcal{F}) = \sup_{f \in \mathcal{F}} E\|\hat{f}_n - f\|_{p'}^{p'},$$

where $\mathcal{F}$ will usually be a subset of densities with fixed compact support and bounded in the norm of one of the Besov spaces $B_{\sigma pq}$. Our main point is that the same form of estimator, based on simple thresholding of the wavelet coefficients, achieves nearly optimal performance, in terms of rates of convergence over a variety of global error measures and over a variety of function spaces. Here, near optimality means that the rates are best possible except possibly for terms logarithmic in sample size. The significance of this universality

of near optimality is discussed in much greater detail in Donoho, Johnstone, Kerkyacharian and Picard (1995).

Concerning the scale of Besov spaces $B_{\sigma pq}$, for the purposes of this section, let us note only that it includes the traditional norms used in statistical theory, namely the Hilbert–Sobolev ($H_2^\sigma = B_{\sigma 22}$) and Hölder ($C^\sigma = B_{\sigma\infty\infty}$, $0 < \sigma \notin \mathcal{N}$). For more general Sobolev spaces, and the interesting special case of functions of bounded total variation, we have the inclusions

$$B_{\sigma p1} \subset H_p^\sigma \subset B_{\sigma p\infty}, \qquad B_{111} \subset \mathrm{TV} \subset B_{11\infty}.$$

Nemirovskii, Polyak and Tsybakov (1985) and Nemirovskii (1985) have shown that, over certain spaces in this scale, no linear estimate can attain even the optimal polynomial rate of convergence. For example, over balls in the total variation norm, and for global L_2 error, the minimax rate among *linear* estimators is $O(n^{-1/2})$, whereas the minimax rate among all estimators is $O(n^{-2/3})$. Thus the Besov scale includes a sufficiently broad range of phenomena to make the near-optimality results for wavelet thresholding estimators interesting. Furthermore, Besov spaces have very specific and interesting properties in functional estimation and approximation theory that we recall in the Appendix.

In Section 3 we investigate the behavior of linear estimators in the context of Besov spaces. We prove, in terms of the rate of convergence of $L_{p'}$-norms, that, when $p' > p$, the linear estimators perceive L_p-smoothness only via Sobolev embedding into an $L_{p'}$-smoothness class corresponding to a lower smoothness and hence leading to a nonoptimal rate of convergence.

Theorem 2 in Section 4 establishes lower bounds for optimal rates of estimation over $B_{\sigma pq}$ among all estimators. Two cases emerge, which we shall call "dense" and "sparse," according as $\varepsilon = \sigma p - (p' - p)/2$ is greater than 0 or less than or equal to 0. These lower bounds are derived by considering perturbations of a fixed density, where the perturbations are combinations of basis functions drawn from an appropriate resolution level. The terms dense and sparse refer to the number of basis functions used to form the perturbation—for example, in the less smooth case when $\varepsilon < 0$, a single basis function is employed. It follows from these lower bounds that when $p' > p$ linear estimators cannot achieve the optimal rate of convergence.

To establish upper bounds for specific wavelet threshold estimators, we use two different approaches. The first (Section 5) consists of a direct evaluation of the $L_{p'}$ losses for $p' \geq p$ over densities in $B_{\sigma pq}$ with support in a fixed interval. Theorem 3 shows that the estimator TW defined using thresholds $\lambda_j = K\sqrt{j/n}$ attains the optimal rate to within logarithmic terms, and attains the exactly optimal rate in the "sparse" case.

Section 6 takes a second approach: approximate the density model by a Gaussian white-noise model and then use results for threshold estimators in the white-noise model derived by Donoho and Johnstone (1996). The Gaussian approximation is done coordinatewise (which is sufficient in the setting of Besov spaces), using the Mallows metric. This approach is at present carried out only for quadratic loss. However, with appropriate choice of thresholds, it

does show that wavelet estimators attain the exactly optimal rate. This is in contrast to the approach using thresholds $\lambda_j = K\sqrt{j/n}$ (Section 4), which in the "dense" case (including quadratic loss), yields rates that are suboptimal by a logarithmic factor. Furthermore, it has been suggested by Nussbaum (1995) that a variant of the white-noise approximation seems likely to be able to deal with more general $L_{p'}$ losses.

This Gaussian approximation technique does not actually simplify the structure of the proof: it rather shifts many of the steps into the Gaussian white-noise setting. In this sense, the direct evaluation method provides an alternative approach; indeed, it could also be used in the white-noise model.

Section 7 concludes with an adaptivity result, Theorem 5, which emphasizes that a single, simple estimator can come within logarithmic terms of optimality simultaneously over a wide range of $L_{p'}$ losses and Besov classes. In fact, one simply uses thresholds $\lambda_j = K\sqrt{j/n}$ over a range

$$n^{1/(1+2r_0)} \le 2^j \le n/\log\, n,$$

where $r_0 + 1$ is the regularity of the wavelet.

Some of the results of this paper were announced without proof in Johnstone, Kerkyacharian and Picard (1992).

2. Besov spaces and wavelets. In this section, we recall definitions and set notation for later use. Some equivalent definitions of Besov spaces, which shed further light on their relevance to density estimation, are reviewed in the Appendix.

2.1. *Multiresolution analysis and wavelets.* Let us recall [cf. Meyer (1990)] that one can construct a function ϕ such that:

1. The sequence $\{\phi(x-k),\ k \in \mathbb{Z}\}$ is an orthonormal family of $L^2(\mathscr{R})$. Let V_0 be the subspace spanned.
2. For all $j \in \mathbb{Z}$, $V_j \subset V_{j+1}$ if V_j denotes the space spanned by $\{\phi_{jk},\ k \in \mathbb{Z}\}$, where $\phi_{jk} = 2^{j/2}\phi(2^j x - k)$.

Then we have $\bigcap_{j\in\mathbb{Z}} V_j = \{0\}$ and, furthermore, if $\phi \in L^2(\mathscr{R})$ and $\int \phi = 1$, $L^2(\mathscr{R}) = \bigcup_{j\in\mathbb{Z}} V_j$ and ϕ is called the multiscale function of the multiresolution analysis $(V_j)_{j\in\mathbb{Z}}$. Various regularity properties can be required of ϕ: we shall here assume that:

3. ϕ is of class $\mathscr{C}^r$, ϕ and every derivative up to order r is rapidly decreasing. In this case, the analysis is said to be *regular*.

In fact, we will assume in succeeding sections that, in addition, ϕ is compactly supported in an interval $[-A,\ +A]$ [e.g., Daubechies' families; see Daubechies (1992)].

Under these conditions, define the space W_j by

$$V_{j+1} = V_j \oplus W_j.$$

There exists a function ψ (the "wavelet") such that:

1. $\{\psi(x-k),\ k\in\mathbb{Z}\}$ is an orthonormal basis of W_0;
2. $\{\psi_{jk},\ k\in\mathbb{Z}, j\in\mathbb{Z}\}$ is an orthonormal basis of $L^2(\mathscr{R})$, where $\psi_{jk} = 2^{j/2}\psi(2^j x-k)$;
3. ψ has the same regularity properties as ϕ.

In addition, we have the decomposition

$$L^2(\mathscr{R}) = V_{j_0}\oplus W_{j_0}\oplus W_{j_0+1}\oplus\cdots.$$

That is, for all $f\in L^2(\mathscr{R})$,

$$f=\sum_{k\in\mathbb{Z}}\alpha_{j_0k}\psi_{j_0k}+\sum_{j\ge j_0}\sum_{k\in\mathbb{Z}}\beta_{jk}\psi_{jk},$$

where

$$\alpha_{jk}=\int f(x)\overline{\phi_{jk}(x)}\,dx,\qquad \beta_{jk}=\int f(x)\overline{\psi_{jk}(x)}\,dx.$$

2.2. *Besov spaces.* We give here the definition of Besov spaces in terms of wavelet coefficients. This is convenient as it gives a description in terms of sequence spaces. In Section A.1 we list a survey of other characterizations of Besov spaces connecting them and explaining their role in approximation theory and nonparametric statistics.

Let ϕ satisfy conditions (1), (2) and (3) with $r>\sigma$, let E be the associated projection operator onto V_j and $D_j = E_{j+1}-E_j$. Besov spaces depend on three parameters $\sigma>0$, $1\le p\le\infty$ and $1\le q\le\infty$ and are denoted $B_{\sigma pq}$. Say that $f\in B_{\sigma pq}$ if and only if the norm

$$J_{\sigma pq}(f)=||E_0(f)||_{L^p(\mathscr{R})}+\left(\sum_{j\ge0}(2^{j\sigma}||D_jf||_{L^p(\mathscr{R})})^q\right)^{1/q}<\infty$$

(with the usual modification for $q=\infty$). Using now the decomposition of f:

$$E_0f=\sum_{k\in\mathbb{Z}}\alpha_{0k}\phi_{0k},$$

$$D_jf=\sum_{k\in\mathbb{Z}}\beta_{jk}\psi_{jk},$$

we may also say that $f\in B_{\sigma pq}$ if and only if the equivalent norm

$$J'_{\sigma pq}(f)=||\alpha_{0\cdot}||_{l_p}+\left(\sum_{j\ge0}(2^{j(\sigma+1/2-1/p)}||\beta_{j\cdot}||_{l_p})^q\right)^{1/q}<\infty$$

[we have set $||\beta_{j\cdot}||_{l_p}=(\sum_{k\in\mathbb{Z}}|\beta_{jk}|^p)^{1/p}$]. Abusing the notation slightly, we will also write $||\beta||_{\sigma pq}$ for the above sequence norm applied to coefficients $((\alpha_{jk}),\ (\beta_{jk}))$. Set also $B_{\sigma pq}(M)=\{\beta\colon\ ||\beta||_{\sigma pq}\le M\}$.

(Note that the L_p Sobolev spaces have a different characterization in terms of sequences [e.g., Frazier, Jawerth and Weiss (1991)].)

This second definition is equivalent to the previous one as a consequence of the following lemma (which will also be useful in the sequel).

LEMMA 1 [Meyer (1990)]. *Let g be such that conditions* (1) *and* (3) *hold. Let $\theta(x) = \theta_g(x) = \sum_{k\in\mathbb{Z}} |g(x-k)|$ and $||\theta||_p = (\int_0^1 |\theta(x)|^p\,dx)^{1/p}$. Let $f(x) = \sum_{k\in\mathbb{Z}} \lambda_k 2^{j/2} g(2^j x - k)$. If $1 \le p \le \infty$ and p_1 satisfies $1/p + 1/p_1 = 1$, then*

$$c_1 2^{j(1/2-1/p)} ||\lambda||_{l_p} \le ||f||_{L_p} \le c_2 2^{j(1/2-1/p)} ||\lambda||_{l_p},$$

where $c_1 = 1/||\theta||_1^{1/p_1} ||\theta||_\infty^{1/p}$ and $c_2 = ||\theta||_p$.

REMARKS.

1. Well-known particular cases of the Besov spaces include the Hilbert–Sobolev spaces $H^\sigma = B_{\sigma 22}$, the set of bounded σ-Lipschitz functions $B_{\sigma\infty\infty}$, $0 < \sigma < 1$, and the Zygmund class $B_{1\infty\infty}$.
2. Using the J or J' norms, the Sobolev embeddings are easily obtained:
$$B_{\sigma' pq'} \subset B_{\sigma pq} \quad \text{for } \sigma' > \sigma \text{ or for } \sigma' = \sigma \text{ and } q' \le q,$$
$$B_{\sigma pq} \subset B_{\sigma' p'q} \quad \text{for } p' > p,\ \sigma' = \sigma - 1/p + 1/p'.$$
In particular, for $\sigma - 1/p > 0$, $q > 1$, $B_{\sigma pq} \subset B_{\sigma'\infty\infty}$ is included in the space of bounded continuous functions. Furthermore, the same is true for $\sigma - 1/p \ge 0$ and $B_{\sigma p1}$.
3. We will also need the inclusion [cf. Meyer (1990) and Peetre (1975)] $B_{0p'p'\wedge 2} \subset L_{p'}$, $p' \ge 1$, where $B_{0p'q}$ is defined through the $J'_{\sigma pq}$ norm by putting $\sigma = 0$.
4. In Section 8.1 we list a survey of other characterizations of Besov spaces, explaining their important role in approximation theory and statistics.
5. The spaces of densities we use are defined by
$$\mathscr{D}_{\sigma pq}(M) = \left\{ f\colon \int f = 1,\ f \ge 0,\ J'_{\sigma pq}(f) \le M \right\},$$
$$\mathscr{D}_{\sigma pq}(M,\ T) = \{ f \in \mathscr{D}_{\sigma pq}(M)\colon \operatorname{supp} f \subset [-T,\ +T] \}.$$

3. Linear estimators. In order to compare the classes of linear and nonlinear estimators, we begin first with the class $\mathscr{C}_L$ of linear estimators, defined by the representation

$$\hat{f}_L(X_1, \ldots, X_n,\ x) = \sum_1^n T_i(X_i,\ x),$$

where $T_i(\cdot,\ \cdot)$ are arbitrary measurable functions. An important class of examples arises as follows. Let $X_1, \ldots, X_n$ be n i.i.d. random variables with common density f and empirical distribution function $F_n = n^{-1} \sum_{i=1}^n I\{X_i \le x\}$. Given a function $E(x,\ y)$, let $E_j(x,\ y) = 2^j E(2^j x,\ 2^j y)$, and consider the linear estimator

$$\hat{E}_{j(n)} = \int E_{j(n)}(x,\ y)\, dF_n(y).$$

Two cases are of particular interest:

$$E^*(x, y) = \kappa(x - y), \tag{4}$$

$$E^{\#}(x, y) = \sum_{k \in \mathbb{Z}} \phi(x - k)\phi(y - k). \tag{5}$$

E^* corresponds to the classical kernel estimate and $E_j^{\#}$ to a projection estimator on the space V_j derived from the scale function ϕ of a multiresolution analysis. Linear estimators have the following distinguishing property. If f, g are two probability densities and $\alpha \in [0, 1]$, then

$$E_{\alpha f + (1-\alpha)g} \hat{f}_L = \alpha E_f \hat{f}_L + (1 - \alpha) E_g \hat{f}_L.$$

The following results will show that the rate of convergence of linear procedures may be strictly slower than that of nonlinear ones. This phenomenon is associated with a difference between the order of integration, p', in the loss function and the order, p, in the regularity constraints. It has already been observed in the related context of regression [Nemirovskii (1985) and Donoho and Johnstone (1996)] and estimation over ℓ_p balls [Donoho and Johnstone (1994)]. In the case of density estimation, we have the following results, beginning first with linear estimators. In the sequel, we denote the risk by $(E_f \|\hat{f}_n - f\|_{p'}^{p'})^{1/p'}$ even in the case $p' = \infty$ where it signifies $E_f \|\hat{f}_n - f\|_\infty$.

In the following result, we shall also need a domination condition in the case $1 \le p' < 2$. Let $\omega \in L^{p/2}(\mathscr{R})$ be an even function, nondecreasing on $\mathscr{R}^+$ and set

$$\mathscr{N}_p = \{f\colon \exists\, a\colon\ f(x - a) \le \omega(x)\}.$$

THEOREM 1. *Let* $1 \le p,\ q \le \infty,\ p' \ge p,\ 1 \le p' < +\infty,\ \sigma > 1/p.$

$$R_n^L = \inf_{\hat{f}_n \in \mathscr{C}_L} \sup_{f \in \mathscr{D}_{\sigma pq}(M)} \left(E_f \|\hat{f}_n - f\|_{p'}^{p'}\right)^{1/p'}.$$

In the case $1 \le p' < 2$, *the set* $\mathscr{D}_{\sigma pq}(M)$ *is replaced by* $\mathscr{D}_{\sigma pq}(M) \cap \mathscr{N}_{p'}$.

There exist constants C_i *such that*

$$C_1 n^{-\sigma'/(1+2\sigma')} \le R_n^L \le C_2 n^{-\sigma'/(1+2\sigma')},$$

where $\sigma' = \sigma - 1/p + 1/p'$.

The same result is true replacing $\mathscr{D}_{\sigma pq}(M)$ *by* $\mathscr{D}_{\sigma pq}(M, T)$ *and* n *by* $n/(\log n)$ *in the case* $p' = \infty$.

PROOF. For the lower bound, we present the details of the proof in the Appendix and give only the idea here. The minimax risk is bounded below by the maximum risk over an ℓ_p ball at a particular resolution level j. For $p' \ge p$, the least favorable points for *linear* estimates over ℓ_p balls are "spikes"—such as the elements of a fixed $\mathscr{P}_j$ as introduced in the proof of Theorem 2 [compare Donoho and Johnstone (1994), Section 8, in the Gaussian case]. The lower bound is obtained by randomizing over the elements of $\mathscr{P}_j$.

For the upper bound, it suffices to exhibit an estimator attaining the right rate of convergence, for example, the "linear wavelet estimator" [cf. Kerkyacharian and Picard (1992)]

$$\hat{f}_{n,j} = \sum_{k\in\mathbb{Z}} \hat{\alpha}_{jk} 2^{j/2} \phi(2^j x - k),$$

where $\hat{\alpha}_{jk} = n^{-1}\sum_{i=1}^{n} \phi_{jk}(X_i)$. We recall that, since ϕ has compact support, the summation in k is finite and ϕ has regularity $r > \sigma$.

PROPOSITION 1. *Suppose* $\pi \geq 1$, $\tau < r$ *and* $f \in \mathscr{D}_{\tau\pi q}(M)$ [*respectively*, $\mathscr{D}_{\tau\pi q}(M) \cap \mathscr{N}_{\pi}$ *if* $1 \leq \pi < 2$ *or* $\mathscr{D}_{\tau\pi q}(\mathrm{M}, T)$ *if* $\pi = \infty$]. *If* $j(n) = [\log_2(n(\log n)^{-I\{\pi=\infty\}})^{1/(1+2\tau)}]$, *there exists a constant* C_3 *such that*

$$\{E_f \|\hat{f}_{n,j(n)} - f\|_{\pi}^{\pi}\}^{1/\pi} \leq C_3 (n(\log n)^{-I\{\pi=\infty\}})^{-\tau/(1+2\tau)}.$$

This result for $\pi < \infty$ is proved in Kerkyacharian and Picard (1992). When f is compactly supported the same argument easily extends to the case $\pi = \infty$ replacing moment bounds by large-deviation inequalities [see (17) below]. The upper bound in Theorem 1 is now a consequence of Proposition 1 and the Sobolev embeddings (see Section 2) $B_{\sigma pq} \subset B_{\sigma' p' q}$ for $p' \geq p$, $\sigma - 1/p = \sigma' - 1/p'$, in which we take $\pi = p'$ and $\tau = \sigma'$.

4. Lower bounds. The corresponding lower bound for *nonlinear* estimators reveals an "elbow" in the rates of convergence. Indeed, let

$$\alpha = \min\left(\frac{\sigma}{1+2\sigma}, \frac{\sigma - 1/p + 1/p'}{1+2\sigma-2/p}\right), \qquad \varepsilon = \sigma p - \frac{p'-p}{2}. \tag{6}$$

We note that

$$\alpha = \begin{cases} \sigma/(1+2\sigma), & \varepsilon \geq 0, \\ \sigma'/(1+2\sigma-2/p), & \varepsilon \leq 0, \end{cases} \tag{7}$$

and also that

$$\begin{aligned} (p'-p)/2 + \varepsilon(1-2\alpha) &= \alpha p' \quad \text{if } \varepsilon \geq 0, \\ (p'-p)/2 + \varepsilon\alpha/\sigma' &= \alpha p' \quad \text{if } \varepsilon \leq 0. \end{aligned} \tag{8}$$

THEOREM 2. *Let* $1 \leq p, q \leq \infty$, $p' \geq p$, $\sigma > 1/p$.

$$R_n = \inf_{\hat{f}} \sup_{f\in\mathscr{D}_{\sigma pq}(M)} \left(E_f \|\hat{f}_n - f\|_{p'}^{p'}\right)^{1/p'}$$

[*the infimum being taken over* all *estimators*, *taking their values in a space containing* $\mathscr{D}_{\sigma pq}(M)$]. *There exists a constant* C_4 *such that*

$$R_n \geq \begin{cases} C_4 \left(\dfrac{\log n}{n}\right)^{\alpha}, & \varepsilon \leq 0, \\ C_4 n^{-\alpha}, & \varepsilon > 0. \end{cases}$$

REMARKS.

1. As will be shown in the next two sections, the lower bound of Theorem 2 is sharp, at least in the cases ($p' \geq p$, $1 < \sigma p < (p' - p)/2$) and ($p = 2$, $\sigma > 1/p$). Also, in all the cases covered by Theorem 3, it is correct up to logarithmic terms.
2. We note two special phenomena. First, an "elbow" appears in the rate of convergence: the "usual" rate ($\sigma/(1+2\sigma)$) applies only if σ is large enough—in other cases, the rate is $\sigma'/(1 + 2\sigma - 2/p)$. Second, a log term appears in the low regularity cases.
3. Comparison with Theorem 1 now shows that linear estimates have suboptimal rates of convergence for $p' > p$ (if also $p' \geq 1$, $\sigma > 1/p$).
4. It is interesting to remark that for $L_{p'}$-loss, linear estimators "perceive" the underlying density L_p-smoothness class via the Sobolev embedding into an $L_{p'}$-smoothness class, with necessarily lower smoothness if $p < p'$ (see Section 2). This phenomenon, which underlies the nonoptimality of linear estimators, has earlier been worked out for $p = 2$ using a theory of "quadratic convexity" [Donoho, Liu and MacGibbon (1990)] and also considering the maximal functional class associated with a fixed minimax rate of convergence for linear estimators [Kerkyacharian and Picard (1993)].

PROOF OF THEOREM 2. We give only a brief sketch, as it is a slight modification of Nemirovskii's method applied to the case of densities. A fuller discussion of lower bounds [including sharper refinements for the case $\varepsilon = 0$ appears in Donoho, Johnstone, Kerkyacharian and Picard (1996)].

For small σ (i.e., $\varepsilon \leq 0$, sparse case), we consider the set of vertices of (one layer of) a pyramid

$$\mathscr{P}_j = \{g_0 \pm \gamma\psi_{jk},\ k \in K_j\} \quad \text{for } j \geq 0,$$

where g_0 is some infinitely differentiable density satisfying $g_0 \geq c$ for x in the interval $[-A, A]$ containing the support of ϕ and ψ. Choose M so that $J'_{\sigma pq}(g_0) \leq M/2$ and let $K_j = \{-(2^j - 1)A + 2lA,\ l = 0, \ldots, (2^j - 1)\}$, so that ψ_{jk} and $\psi_{jk'}$ have disjoint supports for unequal k, $k' \in K_j$. Finally, in order that $\mathscr{P}_j$ be included in $\mathscr{D}_{\sigma pq}(M)$, choose γ such that $0 \leq \gamma \leq \Gamma(j; \sigma, p, M)$, where

$$\Gamma(j; \sigma, p, M) = \frac{C}{||\psi||_\infty} 2^{-j/2} \wedge \frac{M}{2} 2^{-j(\sigma+1/2-1/p)}.$$

The inequality follows by standard arguments using Fano's lemma.

For the case of larger σ (i.e., $\varepsilon \geq 0$, dense case), we consider the set of vertices of a cube

$$C_j = \left\{ f_\varepsilon = g_0 + \sum_{k \in K_j} \gamma\varepsilon_k\, \psi_{jk}, \varepsilon_k = \pm 1 \right\},$$

with

$$0 \le \gamma \le \frac{C2^{-j/2}}{||\psi||_\infty} \wedge \frac{M}{2} 2^{-j(\sigma+1/2)},$$

and using now Assouad's lemma, we obtain the required inequality. □

5. Threshold wavelet estimators. Among nonlinear estimators, we study a very special one: a truncated threshold wavelet estimator. Define empirical coefficients $\hat{\alpha}_{jk}$, $\hat{\beta}_{jk}$ as in (2) and employ hard thresholding:

$$\tilde{\beta}_{jk} = \begin{cases} \hat{\beta}_{jk}, & \text{if } |\hat{\beta}_{jk}| > KC(j)n^{-1/2}, \\ 0, & \text{if } |\hat{\beta}_{jk}| \le KC(j)n^{-1/2}. \end{cases} \tag{9}$$

Then the estimator TW associated with the functions $j_0(n)$, $j_1(n)$, $C(j)$ and K is

$$\mathrm{TW}(x) = \hat{f}_{n,\,j_1} + \hat{D}_{j_1,\,j_0} = \sum_{k\in\mathbb{Z}} \hat{\alpha}_{j_1 k}\phi_{j_1 k}(x) + \sum_{j_1}^{j_0} \sum_{k\in\mathbb{Z}} \tilde{\beta}_{jk}\psi_{jk}(x). \tag{10}$$

Before considering the properties of this estimator, we pause for some motivation. We have seen in preceding sections that the linear wavelet estimator, LW (corresponding to $j_0 < j_1$, and hence no "detail" term $\hat{D}_{j_1,\,j_0}$) cannot be optimal if $p < p'$. This may be explained via the decomposition of the error into bias and variance components. If LW uses level $j(n)$, it has bias of order $2^{-j(n)\sigma' p'}$, while the stochastic term is of order $(2^{j(n)}/n)^{p'/2}$. This leads to the idea of beginning with a low-frequency estimator LW($j_1(n)$), with $j_1(n)$ chosen low enough that the stochastic term has the right rate, and then adding in certain "details" up to the higher order $j_0(n)$ in such a way that the bias term also has the right order. (It is easily seen that, if $p' = p$, it suffices to choose $j_0 < j_1$, whereas, for $p' > p$, it is necessary to take $j_0 > j_1$.)

It remains now to choose a way of refining the details, and this is done using hard thresholding: a superefficiency procedure in the spirit of the Hodges–Lehmann estimator near $\beta_{jk} = 0$. This choice makes sense since the constraint $\mathscr{D}_{\sigma pq}(M)$ on the function "forces" most of the β_{jk} to be small. We focus here on the choice $C(j) = \sqrt{j}$. The first theorem describes the behavior of TW when p, q, σ are known. An adaptivity result for unknown p, q, σ appears in Section 7.

As before, in the proof of Theorem 2 we use the index $\varepsilon = \sigma p - (p' - p)/2$ to distinguish "dense," "critical" and "sparse" cases. In the statement below, the notation $2^{j(n)} \simeq g(n)$ means that $j(n)$ is chosen to satisfy the inequalities $2^{j(n)} \le g(n) < 2^{j(n)+1}$.

THEOREM 3. *Let $\sigma - 1/p > 0$ and $p \wedge 1 \le p' \le \infty$. If $C(j) = \sqrt{j}$, there exist constants $C_5 = C_5(\sigma, p, q, M)$ and $K_0 = K_0(\sigma, p, p'; M)$* [*specified after* (18) *below*] *such that if*

$$\begin{aligned} 2^{j_1(n)} &\simeq \left(n(\log n)^{[(p'-p)/p]I\{\varepsilon\ge 0\}}\right)^{1-2\alpha}, \\ 2^{j_0(n)} &\simeq (n/\log n)^{\alpha/\sigma'} \end{aligned} \tag{11}$$

and $K \geq K_0$, *then*

$$(12)\qquad \sup_{f \in \mathcal{I}_{\sigma pq}(M,T)} (E_f ||\mathrm{TW} - f||_{p'}^{p'})^{1/p'} \leq \begin{cases} C_5 (\log n)^{(1-\varepsilon/\sigma p)\alpha} n^{-\alpha}, & \varepsilon > 0, \\ C_5 (\log n)^{(1/2 - p/qp')_+} \left(\dfrac{\log n}{n}\right)^{\alpha}, & \varepsilon = 0, \\ C_5 \left(\dfrac{\log n}{n}\right)^{\alpha}, & \varepsilon < 0, \end{cases}$$

where $x_+ = \max(x, 0)$.

REMARKS. In the case $\varepsilon < 0$, the rate is sharp: the bounds in Theorems 2 and 3 agree.

When $\varepsilon = 0$, there is an extra logarithmic term when q is sufficiently large. It turns out [Donoho, Johnstone, Kerkyacharian and Picard (1996)] that this extra term is actually sharp, since the lower bound of Theorem 2 can be improved to contain it, at least in the Gaussian white-noise setting. Of course, the constant C_5 depends on p, q, σ, p' and blows up for $\varepsilon \to 0$ or $q \to 2p/p'$, which accounts for the discontinuous nature of the results as presented here.

These logarithmic terms do not appear in the case of quadratic losses $p' = 2$ studied by Donoho and Johnstone (1996) since $\varepsilon \leq 0$ and $\sigma > 1/p$ together imply $p' > 2 + p$.

When $\varepsilon > 0$, the exponent of the logarithmic term in the upper bound is strictly better than $\alpha p'$ and is independent of q [indeed, the simpler choice $2^{j_1(n)} \simeq n^{1-2\alpha}$ leads to the poorer risk bound $C_5(\log n/n)^{\alpha}$]. However, for $\varepsilon > 0$, this logarithmic term is not in fact necessary. For example, it does not appear in the case $p' = 2$ studied by Donoho and Johnstone, and we show in the next section that we can modify $C(j)$ so as to obtain the analog of their result when $p' = 2$. Furthermore, after this manuscript was first drafted, Delyon and Juditsky (1993) showed that the choice $C(j) = \sqrt{j - j_1}$ removed the logarithmic term (essentially by reducing the bias term at the critical resolution level j_1) for general p'. Nevertheless, both these modifications have the disadvantage of strongly depending on the constants p, σ, q, p'. Thus they will not be of use when we want to construct adaptive procedures (see the final section).

The number of levels used is proportional to $\log_2 n$: indeed, $j_1(n) \sim (1 - 2\alpha)\log_2 n$ and $j_0(n) \sim (\alpha/\sigma')\log_2 n$. In particular, we note that $j_1(n) < j_0(n)$ unless $p' = p$, $\varepsilon > 0$, in which case Theorems 1 and 2 show that the linear estimators considered in the previous section are optimal. Thus we will exclude this case from the proof that follows.

The restriction $p' \geq p$ is inessential: for compactly supported functions the L_p-norms are increasing. So in the case $p' < p$ linear estimators attain the bound and there is no need of nonlinear estimates.

5.1. *Proof of Theorem 3.* This section will be divided into two parts. In the first part we set up the technical tools of the proof: moment bounds, large-

deviation results and norm inequalities. In the second part we derive the results, but in outline the argument runs as follows. We consider three terms:

1. a bias term $||f - E_{j_0(n)}f||_{p'}^{p'}$, which will be negligible because of the approximation properties of Besov spaces;
2. the error due to the linear component of the estimator TW,

$$E_f||\hat{f}_{n,\,j_1(n)} - E_{j_1(n)}f||_{p'}^{p'},$$

which will be treated as is usual for kernel estimators, for example, by evaluations of moment bounds;
3. the details term: using norm inequalities, we reduce the treatment of this global term to the specific study of each coefficient $\tilde{\beta}_{jk} - \beta_{jk}$. The approach is then inspired by Hodges–Lehmann superefficiency arguments. A large-deviations approach shows that there is a negligible probability that $\hat{\beta}_{jk}$ and β_{jk} differ greatly. Then only two kinds of errors have to be controlled: $\hat{\beta}_{jk} - \beta_{jk}$, when both of them are large, or β_{jk}, when both are small. We then employ moment bounds in the first case, and for the second term the Besov constraint to show that β_{jk} cannot be large very frequently.

5.1.1. *Preliminaries.*

Moment bounds. We recall the following result of Rosenthal (1972). Let $Y_1, \ldots, Y_n$ be i.i.d. random variables with $EY_i = 0$, $EY_i^2 \le \sigma^2$. Then there exists c_m such that

$$\begin{aligned} &E|n^{-1}\textstyle\sum Y_i|^m \le c_m\left(\frac{\sigma^m}{n^{m/2}} + \frac{E|Y_1|^m}{n^{m-1}}\right) \quad \text{if } m \ge 2, \\ &E|n^{-1}\textstyle\sum Y_i|^m \le \sigma^m n^{-m/2} \quad \text{if } 1 \le m \le 2. \end{aligned} \tag{13}$$

Back in the density estimation setting, let $X_1, \ldots, X_n$ be an i.i.d. sample from a distribution with bounded density f, and let $g \in L_m(\mathscr{R})$ be bounded with $\int g^2 = 1$. Define $g_{jk}(x) = 2^{j/2}g(2^j x - k)$:

$$\gamma_{jk} = \int g_{jk}(x)f(x)\,dx, \qquad \hat{\gamma}_{jk} = n^{-1}\sum_{i=1}^{n} g_{jk}(X_i).$$

Now apply Rosenthal's inequalities to $Y_i = g_{jk}(X_i) - Eg_{jk}(X_1)$. We note that $E|Y_1|^m \le 2^m E|g_{jk}(X_1)|^m \le 2^m \cdot 2^{j(m/2-1)}\|f\|_\infty \|g\|_m^m$ and that

$$\sigma^2 \le \int |g|^2(x-k)f(x/2^j)\,dx \le ||f||_\infty ||g||_2^2 = ||f||_\infty.$$

It follows that there exists a constant c_m depending only on m such that

$$E|\hat{\gamma}_{jk} - \gamma_{jk}|^m \le c_m\left\{||f||_\infty^{m/2} + 2^m||f||_\infty||g||_m^m\left(\frac{2^j}{n}\right)^{(m/2-1)_+}\right\}n^{-m/2}. \tag{14}$$

Now it is easy to show that, if $f \in \mathscr{D}_{\sigma pq}(M)$, then

$$||f||_\infty \le (1-2^{-\sigma'' q'})^{1/q'} J'_{\sigma'' \infty q}(f) \le M(1-2^{-\sigma'' q'})^{1/q'}, \tag{15}$$

where $\sigma'' = \sigma - 1/p > 0$ and $1/q + 1/q' = 1$. Consequently, when $f \in \mathscr{D}_{\sigma pq}(M)$, the bound (14) may be written as

$$E|\hat{\gamma}_{jk} - \gamma_{jk}|^m \le c_{mb} n^{-m/2}, \tag{16}$$

for all j if $1 \le m \le 2$, and as soon as $n \ge 2^j$ for $m > 2$, where c_{mb} depends as shown in (14) and (15) on σ, p, q, M, $||g||_m$ and m.

Large deviations. The terms e_{bs} and e_{sb} below are bounded using large-deviation inequalities for the event $|\hat{\beta}_{jk} - \beta_{jk}| > (K/2)\sqrt{j/n}$. We therefore recall Bernstein's inequality. If $Y_1, \dots, Y_n$ are i.i.d. bounded random variables such that $EY_i = 0$, $EY_i^2 = \sigma^2$, $|Y_i| \le ||Y||_\infty < \infty$, then

$$P\left(\left|n^{-1}\sum Y_i\right| > \lambda\right) \le 2\exp\left(-\frac{n\lambda^2}{2(\sigma^2 + ||Y||_\infty \lambda/3)}\right).$$

Applying this to $Y_i = \psi_{jk}(X_i) - E_f \psi_{jk}(X_1)$ and noting that $\sigma^2 \le ||f||_\infty \le M$, we conclude that, if $j2^j \le n$, then there exists $K = c(M, \psi)\gamma$ such that, for all $\gamma \ge 1$,

$$P\{|\hat{\beta}_{jk} - \beta_{jk}| > (K/2)\sqrt{j/n}\} \le 2^{-\gamma j}. \tag{17}$$

For example, it can be verified that the choice $c(M, \psi) = c\sqrt{M \vee 1}$ suffices if $c = c(\psi)$ is chosen so large that $c^2 \ge 8 \log 2(1 + c\|\psi\|_\infty/3)$.

Norm inequalities. We begin with some useful inequalities for $L_{p'}$-norms ($p' \ge 1$) of a (random) function

$$\hat{f} = \sum_{j_1}^{j_0} \sum_k \hat{f}_{jk} \psi_{jk}.$$

Using the inclusions $B_{0 p' p' \wedge 2} \subset L_{p'}$ and Lemma 1, we have, for $\pi = p' \wedge 2 \ge 1$,

$$||\hat{f}||_{p'}^{p'} \le C^{p'} \left(\sum_{j_1}^{j_0} ||D_j \hat{f}||_{p'}^{\pi} \right)^{p'/\pi}, \tag{18}$$

$$||D_j \hat{f}||_{p'}^{p'} \le C^{p'} 2^{j(p'/2-1)} \sum_k |\hat{f}_{jk}|^{p'}. \tag{19}$$

Here, and throughout, C denotes a constant that is not necessarily the same at each appearance. Define

$$S(\gamma) = \sum_{j_1}^{j_0} 2^{j\gamma} \le \begin{cases} c_\gamma 2^{\max(j_0\gamma,\, j_1\gamma)}, & \gamma \ne 0, \\ (j_0 - j_1), & \gamma = 0. \end{cases} \tag{20}$$

Let $\beta \in \mathscr{R}$ be arbitrary and set $a = p'/(p'-2)$. We will make frequent use of the bound

$$E||\hat{f}||_{p'}^{p'} \leq \begin{cases} C^{p'} \sum_{j_1}^{j_0} 2^{j(p'/2-1)} \sum_k E|\hat{f}_{jk}|^{p'}, & 1 \leq p' \leq 2, \\ C^{p'} S(\beta a)^{(p'/2-1)_+} \sum_{j_1}^{j_0} 2^{j(p'/2-1-\beta p'/2)} \sum_k E|\hat{f}_{jk}|^{p'}, & p' > 2. \end{cases} \tag{21}$$

The first inequality is immediate from (18) and (19). When $p' > 2$, we first apply Hölder's inequality in (18) to obtain, for arbitrary β,

$$\left(\sum ||D_j \hat{f}||_{p'}^2\right)^{p'/2} \leq \left(\sum_{j_1}^{j_0} 2^{j\beta p'/(p'-2)}\right)^{p'/2-1} \sum_{j_1}^{j_0} 2^{-j\beta p'/2} ||D_j \hat{f}||_{p'}^{p'}. \tag{22}$$

Combining (22) with Lemma 1 yields the second inequality in (21). If we adopt the purely formal convention that $S^0 = 1$, then the second inequality in (21) (in the particular case of $\beta = 0$) reduces to the first, and so, with this convention, we use the second inequality of (21) for all $p' \geq 1$ below.

5.1.2. *Completion of the proof.* The estimator TW in (10) has two parts: a linear piece $\hat{f}_{n,\, j_1(n)}$ and a detail term $\hat{D}_{j_1,\, j_0}$. Along with a corresponding decomposition of $f = E_{j_1} f + D_{j_1,\, j_0} f + (f - E_{j_0} f)$, this yields

$$\begin{aligned} E_f||\mathrm{TW} - f||_{p'}^{p'} \leq 3^{p'-1}(&E_f||\hat{f}_{n,\, j_1(n)} - E_{j_1(n)} f||_{p'}^{p'} \\ &+ E_f||\hat{D}_{j_1 j_0} - D_{j_1 j_0} f||_{p'}^{p'} + ||f - E_{j_0(n)} f||_{p'}^{p'}), \end{aligned} \tag{23}$$

where

$$E_j f(x) = \int \sum_{k \in \mathbb{Z}} \phi_{jk}(y) \phi_{jk}(x) f(y)\, dy,$$

$$D_{j_1 j_0} f(x) = \int \sum_{j_1}^{j_0} \sum_{k \in \mathbb{Z}} \psi_{jk}(y) \psi_{jk}(x) f(y)\, dy.$$

The third and first terms in (23) are easily estimated. We start with the approximation error.

We detail the proof only for $p' < +\infty$. The arguments extend easily to the case $p' = +\infty$ if we replace moment bounds evaluations by Bernstein's inequality.

Bias term. Using the characterizations of Besov spaces and the Sobolev embeddings $B_{\sigma pq} \subset B_{\sigma' p' \infty}$, it is easy to see that

$$||f - E_{j_0(n)} f||_{p'}^{p'} \leq C \|f\|_{\sigma pq}^{p'} 2^{-j_0(n)\sigma' p'}. \tag{24}$$

From the choice of $j_0(n)$, this bound has the rate of convergence specified in (12) if $\varepsilon > 0$, $p' = p$; or $\varepsilon = 0$, $p'/2p \leq 1/q$; or $\varepsilon < 0$ and is negligible otherwise.

Linear term. $E_f||\hat{f}_{n,j_1(n)} - E_{j_1}f||_{p'}^{p'}$. Using Lemma 1, (16) and the compact support of ϕ, this term is bounded by

$$||\theta_\phi||_{p'}^{p'} 2^{j_1(n)(p'/2-1)} \sum_{k\in\mathbb{Z}} E|\hat{\alpha}_{j_1(n)k} - \alpha_{j_1(n)k}|^{p'} \le C c_{bh}(T+A)\left(\frac{2^{j_1(n)}}{n}\right)^{p'/2}. \tag{25}$$

From the choice of $j_1(n)$, this bound has the specified rate of convergence if $\varepsilon = 0$, $p'/2p \le 1/q$ and is negligible otherwise.

Details term. To decompose the details term, define

$$\begin{aligned} \hat{B}_j &= \{k\colon |\hat{\beta}_{jk}| > K\sqrt{j/n}\}, & \hat{S}_j &= \hat{B}_j^c, \\ B_j &= \{k\colon |\beta_{jk}| > (K/2)\sqrt{j/n}\}, & S_j &= B_j^c, \\ B'_j &= \{k\colon |\beta_{jk}| > 2K\sqrt{j/n}\}, & S'_j &= B_j'^c. \end{aligned}$$

We may then write

$$\begin{aligned} \hat{D}_{j_1 j_0}f - D_{j_1 j_0}f &= \sum_{j_1}^{j_0}\sum_k (\hat{\beta}_{jk} - \beta_{jk})\psi_{jk}[I\{k\in\hat{B}_j\cap S_j\} + I\{k\in\hat{B}_j\cap B_j\}] \\ &\quad - \sum_{j_1}^{j_0}\sum_k \beta_{jk}\psi_{jk}[I\{k\in\hat{S}_j\cap B'_j\} + I\{k\in\hat{S}_j\cap S'_j\}] \\ &= (e_{bs}+e_{bb}) - (e_{sb}+e_{ss}). \end{aligned}$$

Large-deviation terms. For the term e_{bs}, we set $\hat{f}_{jk} = (\hat{\beta}_{jk} - \beta_{jk})I\{k \in \hat{B}_j S_j\}$. Clearly, $\hat{B}_j S_j \subset D_{jk} = \{|\hat{\beta}_{jk} - \beta_{jk}| > (K/2)\sqrt{j/n}\}$, the large-deviation event studied in (17). We first calculate, using this, Hölder's inequality and (16), that

$$\begin{aligned} \sum_k E|\hat{f}_{jk}|^{p'} &\le \sum_k E\{|\hat{\beta}_{jk} - \beta_{jk}|^{p'}, D_{jk}\} \\ &\le \sum_k (E|\hat{\beta}_{jk} - \beta_{jk}|^{p'r})^{1/r} P(D_{jk})^{1/r'} \\ &\le c_{mb}(T+2A)n^{-p'/2}2^{j(1-\gamma/r')}. \end{aligned}$$

Applying (21) gives

$$E||e_{bs}||_{p'}^{p'} \le C^{p'}\cdot c_{mb} n^{-p'/2}\cdot S(\beta a)^{(p'/2-1)_+} S((1-\beta)p'/2 - \gamma/r'). \tag{26}$$

Using the notation of (20), we note that, when $p' > 2$,

$$S(\beta a)^r S(b) \le c_{p'b\beta} 2^{(\beta p'/2+b)j_s}, \qquad j_s = \begin{cases} j_1, & \text{if } \beta, b < 0, \\ j_0, & \text{if } \beta, b > 0. \end{cases}$$

When $p' \le 2$, we have

$$S(b) \le c2^{bj_s}, \qquad j_s = \begin{cases} j_1, & \text{if } b < 0, \\ j_0, & \text{if } b > 0. \end{cases}$$

Since γ can be chosen arbitrarily large and the choice of β is free (when $p' > 2$), we may arrange that the appropriate arguments of $S(\cdot)$ in (26) (i.e., both when $p' > 2$ and the second when $1 \le p' \le 2$) are negative. Thus, for $p' \ge 1$,

$$E||e_{bs}||_{p'}^{p'} \le C2^{j_1(p'/2-\gamma/r')}n^{-p'/2}.$$

For any choice of $\gamma > 0$, this bound is smaller than the linear term in (25) and so is asymptotically negligible.

For the term e_{sb}, apply (21) and (20) to $\hat{f}_{jk} = \beta_{jk}I\{k \in \hat{S}_j B'_j\}$. Again, $\hat{S}_j B'_j \subset D_{jk}$ and so, using the large-deviation bound and the inclusion $B_{\sigma' p' q} \subset B_{\sigma' p' \infty}$,

$$\begin{aligned} \sum_k E|\hat{f}_{jk}|^{p'} &\le \sum_k |\beta_{jk}|^{p'} P(D_{jk}) \le ||\beta_{j\cdot}||_{p'}^{p'} 2^{-\gamma j} \\ &\le C||f||_{\sigma' p' \infty}^{p'} 2^{-j(\sigma' p' + p'/2 - 1 + \gamma)}. \end{aligned} \tag{27}$$

Thus

$$\begin{aligned} E||e_{sb}||_{p'}^{p'} &\le CS(\beta a)^{(p'/2-1)_+} S(-p'(\beta/2 + \sigma') - \gamma)M^{p'} \\ &\le C2^{-j_1(n)(\gamma + \sigma' p')}, \end{aligned} \tag{28}$$

after choosing β and γ as described for e_{bs} and exploiting the embedding $B_{\sigma p \infty} \subset B_{\sigma' p' \infty}$. This term also is seen to be negligible by taking γ large. For example, the choice $\gamma = \gamma_0 p' = (\alpha/(1-2\alpha) - \sigma')p'$ makes (28) of at most the same order as (24), since $2^{-j_1(n)\alpha/(1-2\alpha)} \le 2^{-j_0(n)\sigma'}$. The constant K_0 in Theorem 3 may then be taken as $c(M, \psi)(\gamma_0 p' \vee 1)$, specified in (17).

Main terms. For the term e_{bb}, apply (21) to $\hat{f}_{jk} = (\hat{\beta}_{jk} - \beta_{jk})I\{k \in \hat{B}_j B_j\}$. In this case, using (16),

$$\begin{aligned} \sum_k E|\hat{f}_{jk}|^{p'} &\le c_{mb} n^{-p'/2} \sum_{k \in B_j} \left| \frac{2\beta_{jk}}{K} \sqrt{\frac{n}{j}} \right|^p \\ &\le C||\beta_j||_p^p \; j^{-p/2} n^{-(p'-p)/2} \\ &\le C||f||_{\sigma p \infty}^p 2^{-j(\sigma + 1/2 - 1/p)p} \, j^{-p/2} n^{-(p'-p)/2}. \end{aligned} \tag{29}$$

In the case $\varepsilon \ne 0$, we bound $j^{-p/2}$ by 1 and use (21) and (20) as before:

$$\begin{aligned} E||e_{bb}||_{p'}^{p'} &\le \frac{CM^p}{n^{(p'-p)/2}} S(\beta a)^{(p'/2-1)_+} S(-\varepsilon - \beta p'/2) \\ &\le \frac{C}{n^{(p'-p)/2}} 2^{\max\{-j_0\varepsilon, -j_1\varepsilon\}}. \end{aligned}$$

Comparison with the statement of Theorem 3 shows that these powers are negligible.

In the case $\varepsilon = 0$ (so that $p' > 2$), set $\beta = 0$ in the bound (22) to obtain

$$E||e_{bb}||_{p'}^{p'} \le CM^p \frac{(j_0 - j_1)^{(p'/2-1)}}{n^{(p'-p)/2}} \sum_{j_1}^{j_0} j^{-p/2} \le CM^p \left(\frac{j_0}{n}\right)^{(p'-p)/2}, \tag{30}$$

since $j_0(n)/j_1(n) \sim p'/(p'-2)$ in the case when $p' > 2$. Thus this term is $O((\log n/n)^{\alpha p'})$ as $n \to \infty$.

Finally, we consider the important case e_{ss}, in which $\hat{f}_{jk} = \beta_{jk} I\{k \in \hat{S}_j S'_j\}$. Using the embedding $B_{0p'p'\wedge 2} \subset L_{p'}$ and the structure of sequence norms, we have

$$\|e_{ss}\|_{p'} \le \|(\{\beta_{jk},\ j_1 \le j \le j_0,\ k \in S'_j\})\|_{0p'p'\wedge 2}.$$

The condition $k \in S'_j$ implies $|\beta_{jk}| \le 2K(j_0/n)^{1/2} = \delta_n$, say. Donoho, Johnstone, Kerkyacharian and Picard (1996, 1995) studied a *modulus of continuity*

$$\Omega^0(\delta; \|\cdot\|, B) = \sup\{\|\beta\|:\ \beta \in B,\ |\beta_{jk}| \le \delta\ \forall\ jk\}.$$

Clearly, we have

$$(E\|e_{ss}\|_{p'}^{p'})^{1/p'} \le \Omega^0(\delta_n; \|\cdot\|_{0p'p'\wedge 2},\ B_{\sigma pq}(M)) = \Omega_n$$

say. In fact, Donoho, Johnstone, Kerkyacharian and Picard (1996) considered a version of Ω^0 in which the number of coefficients n_j at level j was finite (namely 2^j), but it can be seen that in the present case, where $p' > p$, the results are actually unchanged when $n_j = \infty$. From Theorem 3 (Besov modulus) of Donoho, Johnstone, Kerkyacharian and Picard (1996), we read off (noting that $\alpha = r/2$) that

$$\Omega_n \le M^{1-2\alpha} \left(2K\sqrt{\frac{j_0}{n}}\right)^{2\alpha} \left(\log \frac{M}{2K}\sqrt{\frac{n}{j_0}}\right)^{e_C}, \tag{31}$$

where

$$e_C = \begin{cases} 0, & \varepsilon \ne 0, \\ \left(\dfrac{1}{2} - \dfrac{p}{p'q}\right)_+, & \varepsilon = 0. \end{cases}$$

(Here we have used the fact that $\varepsilon = 0$ and $\sigma p > 1 \Rightarrow p' > 2$.) Since $j_0(n) \asymp \log n$, we conclude from this argument that

$$\Omega_n \le CM^{1-2\alpha} \log^{e_C} n \cdot \left(\frac{\log n}{n}\right)^{\alpha}.$$

This bound is sufficiently sharp for our purposes when $\varepsilon \le 0$ (and also for *all* ε for the adaptive result of Section 6).

However, for fixed $(\sigma,\ p,\ q,\ p')$, when $\varepsilon > 0$, the exponent of $\log n$ can be improved to $(1 - \varepsilon/sp)\alpha$ by a more detailed examination of the proof of

Theorem 3 of Donoho, Johnstone, Kerkyacharian and Picard (1996). The key point is that we have adjusted the range of $j \in (j_1(n),\ j_0(n))$ in TW to lie entirely above the least favorable level $j^*(n)$, which satisfies $2^{j^*(n)} \asymp n^{1-2\alpha}$ and leads to the bound (31). So, in fact, $\Omega_n \le 2\|\Omega^*(j)I\{j \ge j_1(n)\}\|_{q'}$, where the quantity $\Omega^*(j)$ is defined in Donoho, Johnstone, Kerkyacharian and Picard (1996) and can be seen to satisfy

$$\Omega^*(j) \le 2^{j\bar{\sigma}'}\delta_n n_{0j}^{1/p'},$$

where $\bar{\sigma}' = 1/2 - 1/p'$, $n_{0j} = (M\delta_n^{-1}2^{-j\bar{\sigma}})^p$ and $\bar{\sigma} = \sigma + 1/2 - 1/p$. Note that n_{0j} is less than 2^j for $j \ge j_1$. [Hence the restriction of $n_j = 2^j$ made in Donoho, Johnstone, Kerkyacharian and Picard (1996) does not affect the results, and so can be applied here to unbounded n_j.] Thus

$$\Omega^*(j) \le C\delta_n^{1-p/p'}2^{-(\bar{\sigma}p-\bar{\sigma}'p')j/p'}.$$

Hence $\Omega_n \le 2\|\Omega^*(j)I\{j \ge j_1\}\|_{q'}$ is dominated by $\Omega^*(j_1)$, and substituting the choice (11) for $2^{j_1(n)}$ yields (for $\varepsilon > 0$) the claimed bound

$$\Omega_n \le CM^{1-2\alpha}(\log n)^{(1-\varepsilon/sp)\alpha}n^{-\alpha}. \qquad \square$$

6. Quadratic loss and Gaussian approximation. We now turn to the specific case of squared error loss, $p' = 2$. In this case, we can exhibit estimators having the exact rate of convergence described by the lower bound of Theorem 2. The approach is via white noise approximation, taking advantage of the results of Donoho and Johnstone (1996). [Nussbaum (1996) has suggested how this approach might be extended to $L_{p'}$-losses for $p' \ne 2$; hopefully, this will be set out in detail elsewhere.]

We begin by recalling the Gaussian white-noise model in sequence space:

$$y_{jk} = \theta_{jk} + \varepsilon z_{jk}, \qquad j = 0,\ 1, \ldots,\ k = 0,\ 1, \ldots, 2^j - 1, \tag{32}$$

where z_{jk} are i.i.d. $N(0,\ 1)$ and $\theta = (\theta_{jk})$ is unknown. Suppose that it is desired to estimate θ with squared error loss $||\hat{\theta} - \theta||_2^2 = \sum(\hat{\theta}_{jk} - \theta_{jk})^2$ and it is known that $\theta \in \Theta_{\sigma pq}(M) = \{\theta:\ ||\theta||_{\sigma pq} \le M\}$, where, in this section,

$$||\theta||_{\sigma pq}^q = \sum_{j\ge 0}(2^{j\tau}||\theta_{j.}||_p)^q \tag{33}$$

and $s = \sigma + 2^{-1} - p^{-1}$, $||\theta_{j.}||_p^p = \sum_{k=0}^{2^j-1}|\theta_{jk}|^p$. The nonlinear minimax risk under squared error loss is defined by

$$R_N(\varepsilon,\ \Theta) = \inf_{\hat{\theta}}\sup_{\theta\in\Theta} E\|\hat{\theta} - \theta\|_2^2.$$

From Donoho and Johnstone (1996), it is known that

$$R_N(\varepsilon,\ \Theta_{\sigma pq}(M)) \sim \gamma(\varepsilon)(M\varepsilon^2)^{2\sigma/(2\sigma+1)}, \tag{34}$$

where $\gamma(\varepsilon) = \gamma(\varepsilon;\ \sigma,\ p,\ q,\ M)$ is a continuous, periodic function of $\log_2\varepsilon$ with period 1.

We recall also that coordinatewise threshold estimators can be chosen to be within a bounded factor of being asymptotically minimax. Define a soft threshold rule $\hat{\theta}^{\lambda}$ by $\{\delta_s(y_{jk},\ \lambda_j),\ j = 0,\ 1, \ldots, k = 0, 1, \ldots, 2^j - 1\}$, where $\delta_s(y, \lambda) = \operatorname{sgn}(y)(|y| - \lambda)_+$. Then Donoho and Johnstone (1996) show that in model (32) there exist absolute constants $A_{\sigma pq}$ such that

$$\inf_{\lambda=(\lambda_j)} \sup_{\Theta_{\sigma pq}(M)} E_\theta ||\hat{\theta}^{\lambda} - \theta||_2^2 \le A_{\sigma pq} R_N(\varepsilon, \Theta_{\sigma pq}(M))(1 + o(1)). \tag{35}$$

THEOREM 4. *Suppose either that* $p \ge 1$ *and* $\sigma > p^{-1}$ *or that* $\sigma = p^{-1}$ *and* $p > 1$. *Then there exist* $\tau = \tau(\sigma,\ p,\ q,\ M)$, $c = c(\sigma,\ p,\ T)$ *and* $C_6 = C_6(\sigma,\ p,\ q,\ M)$ *such that*

$$\inf_{\hat{f}_n} \sup_{\mathscr{D}_{\sigma pq}(M, T)} E_f ||\hat{f}_n - f||_2^2 \le C_6 R_N(\tau n^{-1/2},\ \Theta_{\sigma pq}(cM))(1 + o(1)). \tag{36}$$

Estimators of the form (44) *below attain the bound, for choices of* $j_1,\ j_2$ *and* $\{\lambda_j\}$ *to be described below*.

The following approximation lemma is the basic tool in bounding the density estimation risk by a corresponding white-noise model risk. It is proved in the Appendix.

LEMMA 2. *Let the i.i.d. variables* $Y_1, \ldots, Y_n$ *satisfy* $EY_i = 0$, $EY_i^2 = 1$, $|Y_i| \le M$ *and set* $S_n = \sum_1^n Y_i$. *Then there exist absolute constants* c_1, c_2 *and a standard Gaussian variable* Z *such that, whenever* $M^2 n^{-1} \log^3 n \le c_1$,

$$E(n^{-1/2} S_n - Z)^2 \le c_2 M^2 n^{-1}. \tag{37}$$

The following lemma, also proved in the Appendix, describes a bound on the risk of soft-threshold estimators in the Gaussian white-noise model as the noise variance is increased. This will be used to bound a heteroscedastic model by a homoscedastic one.

LEMMA 3. *Let* E_{β,τ^2} *denote expectation when* $Y \sim N(\beta,\ \tau^2)$. *If* $\tau < \bar{\tau}$, *then*

$$E_{\beta,\tau^2}[\delta_s(Y,\ \lambda) - \beta]^2 \le 2E_{\beta,\bar{\tau}^2}[\delta_s(Y,\ \lambda) - \beta]^2. \tag{38}$$

To apply the lemmas, fix (j, k) and note that $\hat{\beta}_{jk}$ has mean β_{jk} and variance $n^{-1}\tau_{jk}^2$, where $\tau_{jk}^2 = \tau_{jk}^2(f) = \operatorname{Var}_f \psi_{jk}(X)$. We use Lemma 2 to construct an approximation $\hat{\gamma}_{jk}$ having an *exact* Gaussian distribution with the same mean and variance. To this end, let $Y_i = (\psi_{jk}(X_i) - \beta_{jk})/\tau_{jk}$ and note that $|Y_i| \le 2||\psi||_\infty 2^{j/2}/\tau_{jk} = M_{jk}$, say. We construct $\hat{\gamma}_{jk} = \beta_{jk} + n^{-1/2}\tau_{jk} Z_{jk}$ by the following recipe.

First, if $\tau_{jk}^2 \ge 4||\psi||_\infty^2 2^j \log^3 n/c_1 n$, then use Lemma 2 to construct Z_{jk} and note that

$$T_4 = E[\hat{\beta}_{jk} - \hat{\gamma}_{jk}]^2 = n^{-1}\tau_{jk}^2 E[n^{-1/2} S_n - Z_{jk}]^2 \tag{39}$$

$$\le 4||\psi||_\infty^2 c_2 2^j n^{-2}. \tag{40}$$

Second, if $\tau_{jk}^2 < 4||\psi||_\infty^2 2^j \log^3 n/c_1 n$, then choose an independent $Z_{jk} \sim N(0, 1)$ and simply use the inequality

$$T_4 \le 2 \operatorname{Var} \hat{\beta}_{jk} + 2n^{-1}\tau_{jk}^2 = 4n^{-1}\tau_{jk}^2 < 16||\psi||_\infty^2 c_1^{-1} 2^j n^{-2} \log^3 n. \tag{41}$$

In either case, we have, therefore, for all j, k, n,

$$T_4 = E[\hat{\beta}_{jk} - \hat{\gamma}_{jk}]^2 \le c_4 2^j n^{-2} \log^3 n. \tag{42}$$

To apply the Gaussian approximation to $\tilde{\beta}_{jk} = \delta_s(\hat{\beta}_{jk}, \lambda_j)$, we first write

$$[\delta(\hat{\beta}_{jk}, \lambda) - \beta_{jk}]^2 \le 2[\delta(\hat{\beta}_{jk}, \lambda) - \delta(\hat{\gamma}_{jk}, \lambda)]^2 + 2[\delta(\hat{\gamma}_{jk}, \lambda) - \beta_{jk}]^2. \tag{43}$$

We shall use the notation $r(\delta_\lambda, \beta; \tau)$ for the *Gaussian* mean squared error $E[\delta_s(\beta + \tau Z, \lambda) - \beta]^2$ for estimation of β from a single Gaussian observation with mean β and variance τ^2. In addition, the mapping $y \to \delta_s(y, \lambda)$ is a *contraction*: $|\delta_s(y_1, \lambda) - \delta_s(y_2, \lambda)| \le |y_1 - y_2|$ regardless of the value of λ. Thus

$$\begin{aligned} E[\tilde{\beta}_{jk} - \beta_{jk}]^2 &\le 2E[\hat{\beta}_{jk} - \hat{\gamma}_{jk}]^2 + 2r(\delta_\lambda, \beta_{jk}; n^{-1/2}\tau_{jk}) \\ &\le 2c_4 2^j n^{-2} \log^3 n + 4r(\delta_\lambda, \beta_{jk}; \tau n^{-1/2}), \end{aligned}$$

where we have used the approximation error bound (42), the variance bound (38) and τ^2 is any common upper bound on τ_{jk}^2. For example, all densities $f \in \mathscr{F}'_{\sigma pq}(M)$ are uniformly bounded, say by B_0, and so $\tau_{jk}^2 \le \int \psi_{jk}^2(x) f(x)\, dx \le B_0$.

PROOF OF THEOREM 4. It suffices to restrict attention to estimators of the form

$$\hat{f} = \sum_k \hat{\alpha}_{j_1 k} \phi_{j_1 k} + \sum_{j_1}^{j_2} \sum_{k \in \mathbb{Z}} \delta_s(\hat{\beta}_{jk}, \lambda_j) \psi_{jk}. \tag{44}$$

where j_1 is a *fixed* constant and $j_2 = j_2(n)$ will be specified below. Thus

$$\begin{aligned} E||\hat{f} - f||_2^2 &= \sum_k E[\hat{\alpha}_{j_1 k} - \alpha_{j_1 k}]^2 + \sum_{j_1}^{j_2} \sum_k E[\delta_s(\hat{\beta}_{jk}, \lambda_j) - \beta_{jk}]^2 + \sum_{j_2+1}^{\infty} \sum_k \beta_{jk}^2 \\ &= L_n(f) + S_n(f) + T_n(f). \end{aligned}$$

Since j_1 is fixed, $L_n \le Cn^{-1}$ is negligible. A simple maximization shows that

$$\sup\{T_n(f),\ f \in \mathscr{S}_{\sigma pq}(M, T)\} = M^2 2^{-2j_2 s},$$

where $s = \sigma + 1/2 - 1/p$. To bound $S_n(f)$, let $S_j = \{k\colon |2^{-j}k| < T + A\}$, employ (44) and note that

$$\sum_{j_1}^{j_2} \sum_{k \in S_j} 2^j n^{-2} \log^3 n \le 4(T + A) 2^{2j_2} n^{-2} \log^3 n.$$

In summary,

$$E\|\hat{f} - f\|_2^2 \le Cn^{-1} + 4\sum_j \sum_{k\in S_j} r(\delta_{\lambda_j}, \beta_{jk}; \tau n^{-1/2}) + c_5 2^{2j_2} n^{-2}\log^3 n + M^2 2^{-2j_2 s}. \tag{45}$$

Choose j_0 so that $2^{j_0-1} \le T + A < 2^{j_0}$. Using the identification $\theta_{j'k} = \beta_{jk}$, $j' = j + j_0$, and $\bar{\lambda}_{j'} = \lambda_{j+j_0}$, the sum in (45) is bounded by

$$\sup\left\{ \sum_{j'=j_0}^{\infty} \sum_{|k|\le 2^{j'}} r(\delta_{\bar{\lambda}_j}, \theta_{j'k}; \tau n^{-1/2}):\ \theta \in \Theta_{\sigma pq}(2^{j_0 s} M) \right\},$$

which, for appropriate choice of $\bar{\lambda}_j$, is bounded by

$$A_{\sigma pq} R_N(\tau n^{-1/2}, \Theta_{\sigma pq}(2^{j_0 s} M))(1 + o(1)).$$

Thus, for $c = c(\sigma,\ p,\ T)$, we might take $c = 2^s(T + A)^s$.

To complete the proof, it therefore remains to show that the cutoff $j_2 = j_2(n)$ can be chosen so that the final two right-side terms in (45) are of smaller order than R_N, namely, $n^{-2\sigma/(2\sigma+1)}$ [cf (34)]. A sufficient condition for this is easily seen to be

$$\frac{\sigma}{2\sigma+1}\frac{1}{s}\log_2 n \ll j_2(n) \ll \frac{\sigma+1}{2\sigma+1}\log_2 n - \frac{3}{2}\log_2\log_2 n, \tag{46}$$

where $a_n \ll b_n$ is to be interpreted as $b_n - a_n \to \infty$. In turn, a sufficient condition for this is that $\sigma < (\sigma+1)(\sigma+2^{-1}-p^{-1})$, which is certainly satisfied if $p \ge 1$ and either $\sigma = p^{-1} < 1$ or $\sigma > p^{-1}$. □

7. Adaptation results. This section shows that a slight modification of TW renders it adaptive, in the sense that it either exactly or approximately achieves the rates of convergence of Theorem 3 without the need to specify $\sigma,\ p,\ q$. Fix an integer r_0 and define a class

$$\mathscr{S} = \{(\sigma,\ p,\ q,\ T):\ (1/p) < \sigma \le r_0,\ 1 \le q,\ p \le \infty,\ 0 < T < +\infty\}.$$

The modification, denoted ATW, is obtained from compactly supported and $(r_0 + 1)$-regular functions ϕ, ψ in (10) simply by specifying $C(j) = \sqrt{j}$ as before, and

$$2^{j_1(n)} \simeq n^{1/(1+2r_0)}, \qquad 2^{j_0(n)} \simeq n/\log_2 n.$$

The constant K in (9) is chosen as $c(M,\ \psi)r_0 p'$. Thus ATW is constructed from TW by maximizing over $\mathscr{S}$ the range of levels j over which thresholding occurs in (12).

THEOREM 5. *Suppose that $X_1, \dots, X_n$ are i.i.d. with density f of compact support contained in $[-T, T]$ and belonging to some class $\mathscr{D}_{\sigma pq}(M, T)$, where $(\sigma, p, q) \in \mathscr{S}$. If $p' \geq 1$, then, for all $(\sigma, p, q, T) \in \mathscr{S}$, there exists $C_7(\sigma, p, q, M)$ such that*

$$\{E_f||\mathrm{ATW} - f||_{p'}^{p'}\}^{1/p'} \leq \begin{cases} C_7(\log n/n)^{\alpha}, & \varepsilon \neq 0, \\ C_7(\log n)^{(1/2-p/qp')_+}(\log n/n)^{\alpha}, & \varepsilon = 0. \end{cases} \tag{47}$$

REMARK. Although the estimator does not depend on (σ, p, q), its specification still depends on p' and M. A fully adaptive estimator is possible in the Gaussian white-noise case; see Donoho, Johnstone, Kerkyacharian and Picard (1995). An adaptive estimator which also obtains the correct rate of convergence when $\varepsilon > 0$ has recently been constructed by Birgé and Massart (1996) as a byproduct of their complexity penalized model selection approach.

PROOF OF THEOREM 5. Because L_p-norms decrease in p for compactly supported functions, the case $p' < p$ reduces to the case $p' = p$. Thus we investigate only the case $p' \geq p$ and modify the proof of Theorem 3. Here also, we present only the case $p' < \infty$, the extension going through as well. Consider $f \in \mathscr{D}_{\sigma pq}(M)$ and define indices $j_i(\sigma, p, q)$ by

$$2^{j_1(\sigma, p, q)} \simeq (n(\log n)^{-I\{\varepsilon > 0\}})^{1-2\alpha}, \qquad 2^{j_0(\sigma, p, q)} \simeq (n(\log n)^{-I\{\varepsilon \leq 0\}})^{\alpha/\sigma'}.$$

The index $j_1(\sigma, p, q)$ differs only slightly from that used in Theorem 3, which will be denoted $j_1^*(\sigma, p, q)$. Of course, $j_1(n) \leq j_1(\sigma, p, q) \leq j_1^*(\sigma, p, q) \leq j_0(\sigma, p, q) \leq j_0(n)$.

Linear and bias terms. On $\mathscr{D}_{\sigma pq}(M)$, the bias and linear terms have rates of convergence no worse than TW:

$$E_f||E_{j_0}f - f||_{p'}^{p'} \leq C2^{-j_0(n)\sigma' p'} \leq C2^{-j_0(\sigma, p, q)\sigma' p'},$$

$$E_f||\hat{f}_{n, j_1} - E_{j_1}f||_{p'}^{p'} \leq C\left(\frac{2^{j_1(n)}}{n}\right)^{p'/2} \leq C\left(\frac{2^{j_1^*(\sigma, p, q)}}{n}\right)^{p'/2}.$$

Large-deviation terms. The asymptotic behavior of the large-deviation terms e_{sb}, e_{bs} is treated exactly as for TW. Note that $p' \geq p$ implies that $r_0 p' \geq 1$, as required for the large-deviation condition (17). For $\gamma \geq \max(\gamma_0 p', 1)$, with $\gamma_0 = \gamma_0(\sigma, p, p') = (\alpha/(1 - 2\alpha) - \sigma')$, they are bounded by $C2^{-j_0(\sigma, p, q)\sigma' p'}$. In view of the choice $K = c(M, \psi)r_0 p'$, it suffices to verify that $\gamma_0(\sigma, p, p') \leq r_0$ over $\mathscr{S}$. For $\varepsilon > 0$, $\gamma_0 = \sigma - \sigma' \leq r_0$, whereas, for $\varepsilon \leq 0$, $\gamma_0 = 2\sigma'/(p' - 2) \leq 1/p - 1/p' = \sigma - \sigma' \leq r_0$.

Main terms. As noted in the discussion of e_{bb} in Section 4, the argument given there establishes the bounds given in (47) even for the wider range of j considered in ATW.

The behavior of the term e_{bb} is a little more delicate. We look first at the case $\varepsilon \leq 0$, which, as noted earlier, arises only for $p' > 2$. Applying (21) for

$\beta = 0$ with (29) gives

$$E||e_{bb}||_{p'}^{p'} \le C^{p'}(j_0 - j_1)^{(p'-2)/2} c_{mb} n^{-p'/2} \sum_{j_1}^{j_0} 2^{j(p'/2-1)} \sum_k \left| \frac{2\beta_{jk}}{K} \sqrt{\frac{n}{j}} \right|^{p_1}$$

$$\le C^{p'} \left(\frac{j_0}{n} \right)^{(p'-p_1)/2} \left[\sup_j 2^{j(p'-2)/2p_1} ||\beta_j||_{p_1} \right]^{p_1}.$$

Since $\varepsilon \le 0$, (8) shows that $(p' - p)/2 \ge \alpha p'$ so we choose $p_1 \in (p, p')$ so that $(p' - p_1)/2 = \alpha p'$. Thus $p_1 = p'(1 - 2\alpha)$, and setting $\bar{\sigma} = \sigma + 1/2 - 1/p$, it follows that

$$p_1 = p'(1 - \sigma'/\bar{\sigma}) = (p' - 2)/2\bar{\sigma}.$$

Hence $(p' - 2)/2p_1 = \bar{\sigma}$, and since $\|\beta_j\|_p$ increases as p decreases (from p_1 to p), the above supremum is bounded by $\|f\|_{\sigma p \infty}$. Thus

$$E||e_{bb}||_{p'}^{p'} \le C^{p'} \left(\frac{j_0}{n} \right)^{\alpha p'} ||f||_{\sigma p \infty}^{p_1} \le C^{p'} M^{p_1} \left(\frac{\log n}{n} \right)^{\alpha p'}.$$

When $\varepsilon > 0$, we decompose

$$e_{bb} = \left(\sum_{j_1}^{j_1(\sigma pq)} + \sum_{j_1(\sigma pq)}^{j_0} \right) \sum_k (\hat{\beta}_{jk} - \beta_{jk}) \psi_{jk} I\{k \in \hat{B}_j \cap B_j\}$$

$$= e_{bba} + e_{bbb}.$$

The term e_{bbb} is bounded exactly as in the previous section since the upper limit j_0 does not affect the estimate. For the term e_{bba}, we exploit (21) along with (25) (applied to $\hat{\beta}_{jk}$ instead of $\hat{\alpha}_{jk}$) to conclude that

$$E||e_{bba}||_{p'}^{p'} \le C^{p'} S(\beta\alpha)^{(p'/2-1)_+} \sum_{j_1}^{j_1(\sigma pq)} 2^{-j\beta p'/2} c_{mb} (T + A) \left(\frac{2^j}{n} \right)^{p'/2}$$

$$\le C^{p'} \left(\frac{2^{j_1(\sigma pq)}}{n} \right)^{p'/2} \le C^{p'} n^{-\alpha p'}. \qquad \square$$

APPENDIX

A.1. Characterizations of Besov spaces. We list here three further characterizations of Besov spaces. The first explains their role in linear minimax theory, the second their importance in approximation theory. The third is the most usual definition in terms of modulus of continuity.

A.1.1. *Minimax viewpoint.* Let V be a set of densities included in a ball in L_p. We recall the definitions and notation of Section 3 for linear estimators. In particular, let E^l, $l = *, \#$, be the kernels (4) and (5) and let $E_j^l(f) = \int E_j^l(x, y) f(y)\, dy$.

THEOREM 6 [Kerkyacharian and Picard (1993)]. *Let* $2 \leq p \leq \infty$ *and suppose that* V *is a set of densities contained in a ball of* $L_p(\mathscr{R})$ *such that*

1. *There exists* $C_2 > 0$, $\sigma > 0$ *such that, for all* n,

$$\inf_{\hat{f} \in F_n} \sup_{f \in V} E_f ||\hat{f} - f||_p^p \geq C_2 n^{-\sigma p/(1+2\sigma)}, \tag{48}$$

   where F_n *is a set of estimators based on* $X_1, \ldots, X_n$ *containing at least the class of linear estimators*.
2. *There exist a kernel* E^* *with* κ *integrable or* $E^\#$ *with* ϕ *localized and sufficiently smooth and a sequence* $j(n)$ *such that, for* $l = *$ *or* #,

$$\sup_{f \in V} E_f ||\hat{E}^l_{j(n)} - f||_p^p < C n^{-\sigma p/(1+2\sigma)}. \tag{49}$$

Then V *is included in a ball* B *of* $B_{\sigma p \infty}$, *and the problems have the same complexity*: (48) *and* (49) *hold with* V *replaced by* B.

To paraphrase the theorem: sets where linear estimators attain the minimax rate are contained in $B_{\sigma p \infty}$ balls.

A.1.2. *Approximation theory.* The next result is well-known folklore [see Fix and Strang (1969)]: it is also implicit in Peetre (1975); for some details, see Härdle, Kerkyacharian, Picard and Tsybakov (1996).

THEOREM 7. *Let* $N \in \mathscr{N}$, $0 < \sigma < N + 1$, $1 \leq p \leq \infty$, $1 \leq q \leq \infty$. *Suppose that* $\forall\, x, y$, $|E^l(x, y)| \leq H(|x - y|)$ *for* $l = *$ *or* # *with* $H \in L$, *and* $\int H(u)|u|^{N+1}\, du < \infty$:

1. *If* $\int E^l(x, y)(x - y)^k\, dy = \delta_{0,k}$, $\forall\, k = 0, \ldots, N$, *then*

$$\{f \in B_{\sigma pq}\} \quad \textit{implies} \quad \{f \in L_p,\ \varepsilon_j = 2^{j\sigma} ||E^l_j f - f||_p \in l_q(\mathscr{N})\}.$$

2. (a) $\{f \in L_p,\ \varepsilon_j = 2^{j\sigma} ||E^*_j f - f||_p \in l_q(\mathscr{N})\}$ *implies* $\{f \in B_{\sigma pq}\}$.
   (b) *If* ϕ^{N+1} *exists and satisfies* $\sum_{k \in \mathbb{Z}} |\phi^{(N+1)}(x - k)| < M$ *for all* $x \in \mathscr{R}$, *then*

$$\{f \in B_{\sigma pq}\} \Longleftrightarrow \{f \in L_p,\ \varepsilon_j = 2^{j\sigma} ||E^\#_j f - f||_p \in l_q(\mathscr{N})\},$$

   and the norms are equivalent.
3. *In case* $l = \#$, *a sufficient* (*but not necessary condition*) *to assure* $\int E^l(x, y)(x - y)^k\, dy = \delta_{0,k}$, $\forall\, k = 0, \ldots, N$, *is that* ϕ^N *exists in a weak sense and belongs to* L_p *for* $p < +\infty$ (*resp., is uniformly continuous and bounded if* $p = +\infty$).

This characterization in terms of approximation rates is one of the most important properties of Besov spaces. For example, condition $\{f \in L_p,\ \varepsilon_j = 2^{j\sigma} ||E^*_j f - f||_p \in l_q(\mathscr{N})\}$ is necessary but not sufficient for membership in the classical Sobolev spaces.

We note also that part (1) of the above theorem applies to the Haar basis (in the case $N = 0$)—note that it is only the direction in the theorem that is used in our arguments.

A.1.3. *Modulus of continuity* [cf. Bergh and Löfström (1976) and Meyer (1990)]. Suppose that $0 < s < 1$, $1 \le p, q \le \infty$ and set $\tau_h f(x) = f(x-h)$. Set

$$\gamma_{\sigma pq}(f) = \left(\int_{\mathscr{R}} \left(\frac{||\tau_h f - f||_p}{|h|^\sigma} \right)^q \frac{dh}{|h|} \right)^{1/q},$$

$$\gamma_{\sigma p\infty}(f) = \sup_{h \in \mathscr{R}} \frac{||\tau_h f - f||_p}{|h|^\sigma}.$$

In the case $\sigma = 1$, set

$$\gamma_{1pq}(f) = \left(\int_{\mathscr{R}} \left(\frac{||\tau_h f + \tau_{-h} f - 2f||_p}{|h|} \right)^q \frac{dh}{|h|} \right)^{1/q},$$

$$\gamma_{1p\infty}(f) = \sup_{h \in \mathscr{R}} \frac{||\tau_h f + \tau_{-h} f - 2f||_p}{|h|}.$$

For $0 < \sigma \le 1$ and $1 \le p, q \le \infty$, set $B_{\sigma pq} = \{f \in L_p\colon\ \gamma_{\sigma pq} < \infty\}$, equipped with the norm $||f||_{\sigma pq} = ||f||_p + \gamma_{\sigma pq}(f)$. For $\sigma > 1$, set $\sigma = n + \alpha$, with $n \in \mathscr{N}$ and $0 < \alpha \le 1$. Let $f^{(m)}$ denote the mth derivative of f and set $f \in B_{\sigma pq}$ whenever $f^{(m)} \in B_{\alpha pq}$ for all $m \le n$. This space is equipped with the norm

$$||f||_{\sigma pq} = ||f||_p + \sum_{m \le n} \gamma_{\sigma pq}(f^{(m)}).$$

REMARKS.

1. It is easy to see from the definitions that $B_{\sigma\infty 1}$ for $\sigma \ge 0$ and $B_{\sigma\infty q}$ for $\sigma > 0$, $q > 1$ are contained in the space of bounded continuous functions.
2. There are other characterizations of Besov spaces [e.g., Lions–Peetre interpolations of Sobolev spaces or Littlewood–Paley decompositions; cf. Bergh and Löfström (1976), Peetre (1975) and Triebel (1992)] that we will not need here.

A.2. Lower bound for linear estimators. We consider a subclass of densities:

$$\tilde{V}_j = \left\{ g_0 + \sum_{k \in K_j} \lambda_{jk} \psi_{jk},\ \lambda_{jk} \le \Gamma(j;\ \sigma,\ p,\ M) \right\}.$$

Choose $\gamma > 0$ such that $f_k = g_0 + \gamma \psi_{jk}$ and $f'_k = g_0 - \gamma \psi_{jk}$ belong to $\tilde{V}_j$.

LEMMA 4. *Suppose that f_L is such that $E_f \hat{f}_L(x) < \infty$ for all $f \in \tilde{V}_j$ and $x \in \mathscr{R}$. Then*

$$2\gamma \frac{\partial}{\partial \lambda_{jk}} [E_f \hat{f}_L(x)] = E_{f_k} \hat{f}_L(x) - E_{f'_k} \hat{f}_L(x).$$

PROOF.

$$E_{f_k}\hat{f}_L(x) - E_{f'_k}\hat{f}_L(x) = \sum_{i=1}^{n} E_{f_k}T_i(X_i,\ x) - E_{f'_k}T_i(X_i,\ x)$$
$$= 2\gamma \int \sum_{i=1}^{n} T_i(y,\ x)\psi_{jk}(y)\,dy.$$

On the other hand, in $\tilde{V}_j$,

$$E_f\hat{f}_L(x) = \sum_{i=1}^{n} \int T_i(y,\ x)(g_0(y) + \sum_k \lambda_{jk}\psi_{jk}(y))\,dy$$

and

$$\frac{\partial}{\partial \lambda_{jk}} E_f\hat{f}_L(x) = \int \sum_{i=1}^{n} T_i(y,\ x)\psi_{jk}(y)\,dy.$$

This establishes the lemma. □

Let us observe that neither

$$\left(\frac{\partial}{\partial \lambda_{jk}}\right)[E_f\hat{f}_L(x)] \quad \text{nor} \quad a_{jk} := \int \left(\frac{\partial}{\partial \lambda_{jk}}\right)[E_f\hat{f}_L(x)]\psi_{jk}(x)\,dx$$

depends on the choice of $f \in \tilde{V}_j$.

We apply an L_1 version of the Cramér–Rao inequality in the model in which $X_1, \ldots, X_n$ is an i.i.d. sample from $f \in \tilde{V}_j$, $\theta = \lambda_{jk}$ and

$$\hat{T} = \int \hat{f}_L(x)\psi_{jk}(x)\,dx = \hat{\alpha}_{jk}.$$

Indeed,

$$\frac{\partial}{\partial \theta} E_\theta \hat{T} = E_\theta \hat{T}L \le (\sup |L|) \cdot E_\theta|\hat{T}|,$$

where

$$L = \sum_i \frac{1}{f_\theta(x_i)} \frac{\partial}{\partial \theta} f_\theta(x_i) = \sum_i \frac{\psi_{jk}(x_i)}{f_\theta(x_i)}$$

and

$$|L| \le n||\psi||_\infty / C.$$

Thus, for $p' \ge 1$,

$$E_\theta|\hat{T}|^{p'} \ge (E_\theta|\hat{T}|)^{p'}$$
$$\ge C|a_{jk}|^{p'} n^{-p'/2}. \tag{50}$$

Observe now that, if $D_j = E_{j+1} - E_j$ (namely, projection on W_j), then

$$||\hat{f}_L - f||_{p'} \ge a_{p'}||D_j(\hat{f}_L - f)||_{p'}$$

for some constant $a_{p'}$, and hence, from Lemma 1,

$$E_f||\hat{f}_L - f||_{p'}^{p'} \geq a'_{p'} 2^{j(p'/2-1)} E_f \sum_k |\hat{\alpha}_{jk} - \lambda_{jk}|^{p'}. \tag{51}$$

Recalling now the definition of the pyramid $\mathscr{P}_j$ from the proof of Theorem 2,

$$\begin{aligned} R_n(\hat{f}_L) &= \sup_{f \in \mathscr{D}_{\sigma pq}(M)} E_f||\hat{f}_L - f||_{p'}^{p'} \geq \frac{1}{\operatorname{card} \mathscr{P}_j} \sum_{f \in \mathscr{P}_j} E_f||\hat{f}_L - f||_{p'}^{p'} \\ &\geq 2^{-(j+1)} \sum_{k \in K_j} E_{f_k}||\hat{f}_L - f_k||_{p'}^{p'} + E_{f'_k}||\hat{f}_L - f'_k||_{p'}^{p'} \\ &\geq a'_{p'} 2^{j(p'/2-1)} 2^{-(j+1)} \sum_{k \in K_j} \Bigg\{ \sum_{\substack{k' \in K_j \\ k' \neq k}} E_{f_k}|\hat{\alpha}_{jk'}|^{p'} + E_{f'_k}|\hat{\alpha}_{jk'}|^{p'} \\ &\qquad + E_{f_k}|\hat{\alpha}_{jk} - \gamma|^{p'} + E_{f'_k}|\hat{\alpha}_{jk} + \gamma|^{p'} \Bigg\}, \end{aligned} \tag{52}$$

using Lemma 1.

But

$$\begin{aligned} E_{f_k}|\hat{\alpha}_{jk} - \gamma|^{p'} + E_{f'_k}|\hat{\alpha}_{jk} + \gamma|^{p'} &\geq |E_{f_k}\hat{\alpha}_{jk} - \gamma|^{p'} + |E_{f'_k}\hat{\alpha}_{jk} + \gamma|^{p'} \\ &\geq 2^{-(p'-1)}|E_{f_k}\hat{\alpha}_{jk} - E_{f'_k}\hat{\alpha}_{jk} - 2\gamma|^{p'} \\ &= 2\gamma^{p'}|a_{jk} - 1|^{p'}, \end{aligned} \tag{53}$$

using Lemma 4.

Using (50), (52) and (53), we obtain

$$R_n(\hat{f}_L) \geq C a'_{p'} 2^{j(p'/2-2)} \Bigg\{ \sum_{k \in K_j} \gamma^{p'}|a_{jk} - 1|^{p'} + \sum_{k \in K_j} \sum_{\substack{k' \neq k \\ k' \in K_j}} n^{-p'/2}|a_{jk'}|^{p'} \Bigg\}.$$

The double sum collapses to $(2^j - 1)n^{-p'/2} \sum |a_{jk}|^{p'}$, and after setting $\gamma^{p'} = (2^j - 1)n^{-p'/2}$, we have

$$R_n^L \geq a'_{p'} 2^{jp'/2} n^{-p'/2} 2^{-j} \sum_{k \in K_j} \left(|a_{jk} - 1|^{p'} + |a_{jk}|^{p'} \right) \geq c \left(\frac{2^j}{n} \right)^{p'/2}.$$

Recall that γ was constrained to be at most $\Gamma(j; \sigma, p, M)$, which, since $\sigma \geq 1/p$, amounts to requiring that $\gamma \leq (M/2)2^{-j(\sigma+1/2-1/p)}$. To maximize the lower bound subject to this constraint, equate $2^j n^{-p'/2}$ and $2^{-j(\sigma+1/2-1/p)p'}$. This leads to choosing j so that $2^j \asymp n^{1/(1+2\sigma')}$, where $\sigma' = \sigma - 1/p + 1/p'$. It follows that

$$\left(\frac{2^j}{n} \right)^{p'/2} \asymp n^{-\sigma' p'/(1+2\sigma')},$$

which establishes the first part of Theorem 1. □

A.3. Gaussian approximation for quadratic loss.

PROOF OF LEMMA 3. Let us first note some easily verified properties of soft thresholding:

(a) for $\beta, z > 0$, $|\delta_s(\beta - z, \lambda) - \beta| \geq |\delta_s(\beta + z, \lambda) - \beta|$;

(b) for $\beta > 0$, $z \to |\delta_s(\beta - z, \lambda) - \beta|$ is increasing for $0 \leq z < \infty$.

That is, negative disturbances yield bigger errors than positive disturbances of the same size, and the error is monotone in the size of a negative disturbance.

Write $X = \beta + \tau Z$ for $Z \sim N(0, 1)$ and drop explicit reference to λ and s. We now apply these properties in turn:

$$\begin{aligned} E_{\beta,\tau}[\delta(X) - \beta]^2 &= E\{(\delta(\beta + \tau Z) - \beta)^2,\ Z < 0\} \\ &\quad + E\{(\delta(\beta + \tau Z) - \beta)^2,\ Z \geq 0\} \\ &\leq 2E\{(\delta(\beta + \tau Z) - \beta)^2,\ Z < 0\} \\ &\leq 2E\{(\delta(\beta + \bar{\tau} Z) - \beta)^2,\ Z < 0\} \\ &\leq 2E_{\beta,\bar{\tau}}[\delta(X) - \beta]^2. \end{aligned}$$

□

REMARK. Although the constant 2 in the statement of the lemma is not sharp, it cannot be reduced to 1, as may be checked by explicit calculation with $\beta = \lambda$ and τ varying from 0 to ∞.

PROOF OF LEMMA 2. We adopt the following conventions: the notation $x_n = y_n + \theta r_n$ means $|x_n - y_n| \leq r_n$; that is, $\theta \in \mathscr{C}$ satisfies $|\theta| \leq 1$ and may differ at each occurrence. Second, $c_1, c_2, \ldots$ denote absolute constants.

(a) It suffices to assume that the distribution function of the X_i is absolutely continuous. If not, let U_i be i.i.d. uniform and independent of $\{X_i\}$ such that $EU_i = 0$, $EU_i^2 = 1$. The variables $Y_i = X_i \cos\alpha + U_i \sin\alpha$ have absolutely continuous distributions with mean 0, variance 1 and bound $M(1 + \alpha)$. Construct Z by applying the proposition to $S_n^1 = \sum_1^n Y_i$. Since $E(S_n^1 - S_n)^2 \leq n\alpha^2$, the choice $\alpha = n^{-1/2}$ ensures that $E(n^{-1/2}S_n - Z)^2 \leq 2\alpha^2 + 2c_2M^2(1 + \alpha)^2 n^{-1} \leq c_3M^2n^{-1}$.

(b) Let F_n denote the distribution of $W_n = S_n/\sqrt{n}$. Since this is absolutely continuous, the quantile transformation $Z = \Phi^{-1}(F_n(W_n))$ yields a standard Gaussian variable [here Φ denotes the distribution function of a $N(0, 1)$ variate]. We show that Z has the desired approximation by considering in turn large, moderate and small deviations, defined respectively by sets $A_1 = \{w\colon |w| > \sqrt{a \log n}\}$, $A_2 = \{1 \leq |w| \leq \sqrt{a \log n}\}$ and $A_3 = \{|w| \leq 1\}$. Indeed, we write

$$\begin{aligned} E(W_n - Z)^2 &= E\{(W_n - Z)^2,\ |W_n| > \sqrt{a \log n}\} \\ &\quad + \int_{A_2 \cup A_3} [w - \Phi^{-1}(F_n(w))]^2 F_n(dw) \\ &= I_1 + I_2 + I_3. \end{aligned} \tag{54}$$

(c) Small deviations are easily handled by the Berry–Esseen theorem, which implies that $|r_n(x)| = |F_n(x) - \Phi(x)| \le C\rho n^{-1/2} \le CMn^{-1/2}$ since $\rho = E|X_1|^3/(EX_1^2)^{3/2} \le M$. According to the mean value theorem,

$$I_3 \le \int_{-1}^{1} \frac{r_n^2(w)}{\phi^2(u^\star(w))} F_n(dw) \le c_3 M^2 n^{-1}, \tag{55}$$

since $u^\star(w)$ lies between w and $\Phi^{-1}(F_n(w))$, and the latter is bracketed by $\Phi^{-1}(\Phi(w) \pm CMn^{-1/2})$, which in turn is bounded by an absolute constant in view of the assumption on $M^2 n^{-1} \log^3 n \le c_1$.

(d) For large deviations, first use the Hölder inequality to write

$$I_1 \le c_4(E|W_n|^3 + E|Z|^3)^{2/3}\, P^{1/3}\{|W_n| > \sqrt{a \log n}\}. \tag{56}$$

Now use Bennett's inequality [see, e.g., Pollard (1984)] to bound

$$P\left(|S_n| > \sqrt{an \log n}\right) \le 2\exp\{-(1/2)a \log n B(M\sqrt{an^{-1} \log n})\}, \tag{57}$$

where the function $B(\lambda) = 2\lambda^{-2}[(1 + \lambda)\log(1 + \lambda) - \lambda]$ is continuous and decreasing on $[0, \infty]$ with $B(0+) = 1$. By hypothesis, $M^2 n^{-1} \log n \le c_1$, and so the right side is bounded by

$$2\exp\left\{-\frac{a}{2} B(\sqrt{c_1 a}) \log n\right\} \le 2n^{-3} \tag{58}$$

as long as we choose a large ($= 10$ say) and c_1 small enough that $aB(\sqrt{c_1 a}) \ge 6$. Finally, the Rosenthal bound (13) shows that

$$E|W_n|^3 \le c_5(1 + Mn^{-1/2}) \le c_5(1 + c_1^{1/2}), \tag{59}$$

and hence that $I_1 \le c_4(c_5 + E|Z|^3)^{2/3}\, 2^{1/3}\, n^{-1} = c_6\, n^{-1}$.

(e) For moderate deviations, it is sufficient, because of symmetry, to focus on

$$I_2^+ = \int_1^{(a \log n)^{1/2}} [x - \tilde{\Phi}^{-1}(\tilde{F}_n(x))]^2 F_n(dx), \tag{60}$$

where $\tilde{\Phi} = 1 - \Phi$, $\tilde{F}_n = 1 - F_n$. We exploit the following lemma, whose proof we omit.

LEMMA 5. *If $x \ge 1$ and $|\tilde{F}/\tilde{\Phi}(x) - 1| \le e^{-3/2}$, then*

$$|x - \tilde{\Phi}^{-1}(\tilde{F}(x))| \le x^{-1} e^{3/2} |(\tilde{F}/\tilde{\Phi})(x) - 1|. \tag{61}$$

We also use a uniform version of the classical moderate-deviation bound based on the Cramér series [cf. Feller (1971) and Petrov (1975)]. The version we use, due to Sakhanenko (1991), does not require explicit knowledge of the Cramér series $\gamma(x)$. It is phrased instead in terms of the Liapounov exponent $L(h) = \sum_1^n E|Y_i|^3 \max(e^{hY_i}, 1)$, which may be conveniently bounded in our application.

PROPOSITION 2 [Sakhanenko (1991)]. *Let* $W_n = \sum_1^n Y_i$ *be the sum of independent, mean-zero random variables,* $\operatorname{Var} W_n = 1$. *Let* $x \geq 0$ *and* $\tilde{F}_n = P(W_n \geq x)$. *If*

$$16xL(2x) \leq 1, \tag{62}$$

then the Cramér series $\gamma(x)$ *is well defined and satisfies*

$$|\gamma(x)| \leq x^3 L(2x), \tag{63}$$

$$|e^{-\gamma(x)}\tilde{F}_n(x) - \tilde{\Phi}(x)| \leq 32L(2x)\phi(x). \tag{64}$$

In our application, $Y_i = X_i/\sqrt{n}$ are bounded by $Mn^{-1/2}$ and so

$$L(h) \leq Mn^{-1/2}e^{hMn^{-1/2}}. \tag{65}$$

The restriction $1 \leq x \leq \sqrt{a \log n}$ implies that $Mx^3n^{-1/2}$ and hence $Mxn^{-1/2}$ are both bounded by $(a^3M^2n^{-1}\log^3 n)^{1/2} \leq 10^{3/2}\sqrt{c_1}$. For a sufficiently small choice of c_1, we may ensure that $|Mx^3n^{-1/2}| \leq 1/18$, say, and hence that condition (62) holds.

Let $R = \tilde{F}_n(x)/\tilde{\Phi}(x)$ and $\gamma = \gamma(x)$; we exploit the bound

$$|R - 1| \leq e^{\gamma}|e^{-\gamma}R - 1| + |e^{\gamma} - 1|.$$

Combining (63) with (65), we conclude that

$$|\gamma| \leq Mx^3e^{2Mxn^{-1/2}} \leq (1/18)e^{1/9} \leq 1/16.$$

From (64), we obtain

$$|R - 1| \leq 32e^{17/16}L(2x)\phi(x)/\tilde{\Phi}(x) + 2|\gamma(x)|.$$

The function $\nu(x) = \phi(x)/x\tilde{\Phi}(x)$ is decreasing in $x \geq 0$, and so is bounded below in our case by $\nu(1)$. Combining this with (63) again yields, for $1 \leq x \leq \sqrt{a \log n}$,

$$\begin{aligned} |\tilde{F}_n/\tilde{\Phi}(x) - 1| &\leq c_3(x + x^3)L(2x) \leq c_4x^3Mn^{-1/2}e^{2xMn^{-1/2}} \\ &\leq c_5x^3Mn^{-1/2} \\ &\leq c_5 10^{3/2}\sqrt{c_1} \leq e^{-3/2}, \end{aligned} \tag{66}$$

again if c_1 is chosen sufficiently small.

Thus Lemma 5 applies also and, from (66),

$$\begin{aligned} I_2^+ &\leq e^3 \int_1^{(a \log n)^{1/2}} x^{-2}|(\tilde{F}_n/\tilde{\Phi})(x) - 1|^2 F_n(dx) \\ &\leq c_{11}EW_n^4M^2n^{-1} \leq c_{11}(1 + c_1)M^2n^{-1}, \end{aligned} \tag{67}$$

since $EW_n^4 = n^{-2}ES_n^2 \le 1 + M^2 n^{-1} \le 1 + c_1$. This yields the desired bound for I_2^+ and completes the proof of Lemma 2. □

Acknowledgments. We thank Professors Vladimir Zolotarev and Alexandr Sakhanenko for helpful disussions and references to the work on Berry–Esseen theorems used in Section 6. In addition, the referee provided very many helpful suggestions.

The second author would like to thank Université de Paris-Sud (Orsay) for supporting a visit by IMJ.

REFERENCES

BERGH, J. and LÖFSTRÖM, J. (1976). *Interpolation Spaces—An Introduction*. Springer, New York.

BIRGÉ, L. and MASSART, P. (1996). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam* (D. Pollard and G. Yang, eds.). Springer, New York.

DAUBECHIES, I. (1992). *Ten Lectures on Wavelets.* SIAM, Philadelphia.

DELYON, B. and JUDITSKY, A. (1993). Wavelet estimators, global error measures: revisited. Technical Report 782, Institut de Recherche en Informatique et Systèmes Aléatoires, Campus de Beaulieu.

DEVROYE, L. and GYÖRF, L. (1985). *Nonparametric Density Estimation, The L_1 View.* Wiley, New York.

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Minimax risk over ℓ_p-balls for ℓ_q-error. *Probab. Theory Related Fields* **99** 277–303.

DONOHO, D. L. and JOHNSTONE, I. M. (1996). Minimax estimation via wavelet shrinkage. Unpublished manuscript.

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1996). Universal near minimaxity of wavelet shrinkage. In *Festschrift for Lucien Le Cam* (D. Pollard and G. Yang, eds.). Springer, New York.

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: asymptopia? (with discussion). *J. Roy. Statist. Soc. Ser. B* **57** 301–369.

DONOHO, D. L., LIU, R. C. and MACGIBBON, K. B. (1990). Minimax risk over hyperrectangles, and implications. *Ann. Statist.* **18** 1416–1437.

DOUKHAN, P. and LEON, J. (1990). Déviation quadratique d'estimateurs d'une densité par projection orthogonale. *C. R. Acad. Sci. Paris Sér. I Math.* **310** 425–430.

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications* **2**. Wiley, New York.

FIX, G. and STRANG, G. (1969). A Fourier analysis of the finite element method. *Stud. Appl. Math.* **48** 265–273.

FRAZIER, M., JAWERTH, B. and WEISS, G. (1991). *Littlewood–Paley Theory and the Study of Function Spaces*. Amer. Math. Soc., Providence, RI.

HÄRDLE, W., KERKYACHARIAN, G., PICARD, D. and TSYBAKOV, A. (1996). Wavelets and econometric applications. Technical report, Humboldt Univ., Berlin.

JOHNSTONE, I., KERKYACHARIAN, G. and PICARD, D. (1992). Estimation d'une densité de probabilité par méthode d'ondelettes. *C. R. Acad. Sci. Paris Sér. I Math.* **315** 211–216.

KERKYACHARIAN, G. and PICARD, D. (1992). Density estimation in Besov spaces. *Statist. Probab. Lett.* **13** 15–24.

KERKYACHARIAN, G. and PICARD, D. (1993). Density estimation by kernel and wavelet methods: optimality of Besov spaces. *Statist. Probab. Lett.* **18** 327–336.

MEYER, Y. (1990). *Ondelettes et Opérateurs, I: Ondelettes, II: Opérateurs de Calderón–Zygmund, III:* (with R. Coifman)*, Opérateurs Multilinéaires*. Hermann, Paris. (English translation of first volume published by Cambridge Univ. Press.)

NEMIROVSKII, A. (1985). Nonparametric estimation of smooth regression function. *Izv. Akad. Nauk. SSSR Tekhn. Kibernet.* **3** 50–60 (in Russian); *Soviet J. Comput. Systems Sci.* **23** 1–11 (1986) (in English).

NEMIROVSKII, A., POLYAK, B. and TSYBAKOV, A. (1985). Rate of convergence of nonparametric estimates of maximum-likelihood type. *Problems Inform. Transmission* **21** 258–272.
NUSSBAUM, M. (1995). Personal communication.
PEETRE, J. (1975). *New Thoughts on Besov Spaces*. Dept. Mathematics, Duke Univ.
PETROV, V. V. (1975). *Sums of Independent Random Variables*. Springer, New York.
POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.
ROSENTHAL, H. P. (1972). On the span in l_p of sequences of independent random variables. *Israel J. Math.* **8** 273–303.
SAKHANENKO, A. I. (1991). Berry–Esseen type estimates for large deviation probabilities. *Siberian Math. J.* **32** 647–656.
SCOTT, D. (1992). *Multivariate Density Estimation*. Wiley, New York.
SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.
TRIEBEL, H. (1992). *Theory of Function Spaces* **2**. Birkhäuser, Basel.
WALTER, G. (1992). Approximation of the delta function by wavelets. *J. Approx. Theory* **71** 329–343.

DAVID L. DONOHO
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-4065

IAIN M. JOHNSTONE
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-4065

GÉRARD KERKYACHARIAN
URA CNRS 1321
MATHÉMATIQUES ET INFORMATIQUES
UNIVERSITÉ DE PICARDIE
80039 AMIENS
FRANCE

DOMINIQUE PICARD
URA CNRS 1321
MATHÉMATIQUES
UNIVERSITÉ DE PARIS VII
2 PLACE JUSSIEU
75221 PARIS CEDEX 05
FRANCE