Density estimation by wavelet thresholding

David L. Donoho[1],
Iain M. Johnstone[1],
Gérard Kerkyacharian[2],
Dominique Picard[3]

April 1993

Abstract

Density estimation is a commonly used test case for non-parametric 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 $B_{s,p,q}$ and for a range of global L'_p error measures, $1 \leq p' < \infty$. 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 non-linearity 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 ($p' = 2$).

Key Words and Phrases: Minimax Estimation, Adaptive Estimation, Density Estimation, Spatial Adaptation, Wavelet Orthonormal bases, Besov Spaces.

Acknowledgements: We thank Alexandr Sakhanenko for helpful discussions and references to his work on Berry Esseen theorems used in Section 5. This work was supported in part by NSF DMS 92-09130. The second author would like to thank Université de Paris-Sud (Orsay) for supporting a visit by IMJ.

[1] Statistics, Stanford University, Stanford CA, 94305, USA.
[2] Mathématiques et Informatiques, Université de Picardie, Amiens, 80039 France.
[3] Mathématiques, Université de Paris VII, 2 Place Jussieu, 75221 France.

1 Introduction

The recent appearance of explicit orthonormal bases based on multiresolution analyses has exciting implications for non-parametric function estimation. Unlike the traditional Fourier bases, wavelet bases offer a degree of localisation 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 in the context of probability density estimation: estimate a probability density function $f(x)$ on the basis of $X_1, ..., X_n$, independent and identically distributed observations drawn from f. Because of its simple specification, this important practical problem has also served as one of the basic test situations for the theory of non-parametric estimation. An overview of traditional methods and a part of the vast literature on theory and application of density estimation is given by Devroye(1985), Silverman(1986) and Scott (1992). The first use of wavelet bases for density estimation appears in papers by Doukhan and Léon (1990), Kerkyacharian and Picard (1992) and Walter (1990).

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

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

The key advantages of wavelet estimators follow from the effects of even very simple non-linearities involving co-ordinatewise thresholding:

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

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

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

along with $\hat{\alpha}_{j_1 k}$ in (1). 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 have slightly better mean (square) error properties (at the level of constants, not rates), 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

$$R_n(\hat{f}, f) = E||\hat{f}_n - f||_{p'}^{p'} = E\int |f_n(x) - f(x)|^{p'}dx.$$

For the most part, we consider $1 \leq p' < \infty$, 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_{spq}. 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 [DJKP].

Concerning the scale of Besov spaces B_{spq}, for the purposes of this introduction, let us note only that it includes the traditional norms used in statistical theory, namely the Hilbert-Sobolev ($H_2^s = B_{s,2,2}$) and Hölder ($C^\alpha = B_{\alpha,\infty,\infty}, 0 < \alpha \notin \mathcal{N}$). For more general Sobolev spaces, and the interesting special case of functions of bounded total variation, we have the inclusions

$$B_{s,p,1} \subset H_p^s \subset B_{s,p,\infty} \quad , \quad B_{1,1,1} \subset TV \subset B_{1,1,\infty}.$$

Nemirovskii, Tsybakov and Polyak (1984) 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 esimators interesting.

Theorem 2 establishes lower bounds for optimal rates of estimation over $B_{s,p,q}$. Two cases emerge, which we shall call "dense" and "sparse", according as $\epsilon = sp - (p' - p)/2$ is > 0 or ≤ 0. The 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 $\epsilon < 0$, 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 consists of a direct evaluation of the $L_{p'}$ losses for $p' \geq p$ over densities in $B_{s,p,q}$ 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.

A second approach is based on approximating the density model by a Gaussian white noise model and then using results for threshold estimators in the white noise model derived by Donoho and Johnstone (1992). This approach is at present carried out only for quadratic loss but with appropriate choice of thresholds, it can be used to show that wavelet estimators attain the exactly optimal rate. This is perhaps of interest, since the use of quadratic

loss implies that one is in the "dense" case, for which the first approach does not yield exactly optimal rates.

The paper concludes with an adaptivity result, Theorem 4, which emphasises 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

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

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 [M]) that one can construct a function ϕ such that:

(1) the sequence $\{\phi(x-k), k \in \mathcal{Z}\}$ is an orthonormal family of $L^2(\mathcal{R})$. Let V_0 be the subspace spanned.

(2) $\forall j \in \mathcal{Z}$, $V_j \subset V_{j+1}$ if V_j denotes the space spanned by $\{\phi_{jk}, k \in \mathcal{Z}\}$, where $\phi_{jk} = 2^{j/2}\phi(2^j x - k)$.

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

(3) ϕ is of class $\mathcal{C}^r$, 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 [D]).

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) $\{\psi(x-k), k \in \mathcal{Z}\}$ is an orthonormal basis of W_0.

(2) $\{\psi_{jk}, k \in \mathcal{Z}, j \in \mathcal{Z}\}$ is an orthonormal basis of $L^2(\mathcal{R})$, 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

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

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

$$f = \sum_{k\in\mathcal{Z}} \alpha_{j_0 k}\psi_{j_0 k} + \sum_{j\geq j_0}\sum_{k\in\mathcal{Z}} \beta_{jk}\psi_{jk},$$

where

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

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.

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

$$J_{spq}(f) = ||E_0(f)||_{L^p(\mathcal{R})} + (\sum_{j\geq 0}(2^{js}||D_j f||_{L^p(\mathcal{R})})^q)^{1/q} \ < \ \infty.$$

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

$$\begin{aligned} E_0 f &= \sum_{k\in\mathcal{Z}} \alpha_{0k}\phi_{0k} \\ D_j f &= \sum_{k\in\mathcal{Z}} \beta_{jk}\psi_{jk} \end{aligned}$$

we may also say that $f \in B_{spq}$ if and only if

$$J'_{spq}(f) = ||\alpha_{0.}||_{l_p} + \left(\sum_{j\geq 0}(2^{j(s+1/2-1/p)}||\beta_{j.}||_{l_p})^q\right)^{1/q} \ < \ \infty.$$

(we have set $||\beta_{j.}||_{l_p} = (\sum_{k\in\mathcal{Z}} |\beta_{jk}|^p)^{1/p}$).

Note that Sobolev spaces have a different characterisation in terms of sequences (e.g. [FJ]).

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

$$\frac{1}{||\theta||_1^{1/p_1}||\theta||_\infty^{1/p}} 2^{j(1/2-1/p)}||\lambda||_{l_p} \leq ||f||_{L_p} \leq ||\theta||_p 2^{j(1/2-1/p)}||\lambda||_{l_p}$$

Remarks. 1. Using the J or J' norms, the Sobolev embeddings are easily obtained:

$$\begin{aligned} B_{s'pq'} &\subset B_{spq} \text{ for } s' > s, \text{ or } s' = s \text{ and } q' \leq q. \\ B_{spq} &\subset B_{s'p'q} \text{ for } p' > p, s' = s - 1/p + 1/p', \end{aligned}$$

In particular, for $s - 1/p \geq 0$, B_{sp1} is included in the space of bounded continuous functions, and the same is true for $s - 1/p > 0$ and $B_{spq}, q > 1$.

2. Well known particular cases of the Besov spaces include the Hilbertian Sobolev spaces $H^s = B_{s22}$, the set of bounded s-Lipschitz functions $= B_{s\infty\infty}$, and the Zygmund class $B_{1\infty\infty}$.

3. We will also need the inclusion (cf. [M], [P,p. 80]): $B_{0,p',p'\wedge 2} \subset L_{p'}, p' \geq 1$, where $B_{0p'q}$ is defined through the J'_{spq} norm by putting $s = 0$.

4. The spaces of densities we use are defined by

$$F_{spq}(M) = \{f : \int f = 1, f \geq 0, J'_{spq}(f) \leq M\}.$$

3 Linear estimators

In order to compare the classes of linear and non-linear estimators, we begin first with the class $\mathcal{C}_L$ of linear estimators, defined by the representation

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

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 $F_n = n^{-1} \sum_{i=1}^n I\{X_i \leq x\}$.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:

$$E^1(x, y) = \alpha(x - y) \tag{4}$$

$$E^2(x, y) = \sum_{k \in \mathcal{Z}} \phi(x - k)\phi(y - k). \tag{5}$$

E^1 corresponds to the classical kernel estimate and E^2 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 non-linear 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 ([N], [DJ2]), and estimation over ℓ_p- balls [DJ90]. In the case of density estimation, we have the following results, beginning first with linear estimators.

Theorem 1 *Let $1 \le p, q \le \infty$, $p' \ge p$, $s > 1/p$.*

$$R_n^L = \inf_{\hat{f}_n \in \mathcal{C}_L} \sup_{f \in F_{spq}(M)} E_f ||\hat{f}_n - f||_{p'}^{p'}.$$

There exist constants C_i such that

$$C_1 n^{-\frac{s'p'}{1+2s'}} \le R_n^L \le C_2 n^{-\frac{s'p'}{1+2s'}},$$

where $s' = s - 1/p + 1/p'$.

The corresponding lower bound for *non-linear* estimators reveals an 'elbow' in the rates of convergence. Let

$$\alpha = \min(\frac{s}{1+2s}, \frac{s-1/p+1/p'}{1+2s-2/p}) \quad ; \quad \epsilon = sp - \frac{p'-p}{2}. \tag{6}$$

We note that

$$\alpha = \begin{cases} s/(1+2s) & \epsilon \ge 0 \\ s'/(1+2s-2/p) & \epsilon \le 0. \end{cases} \tag{7}$$

Theorem 2 *Let $1 \le p, q \le \infty$, $p' \ge p$, $s > 1/p$.*

$$R_n = \inf_{\hat{f}} \sup_{f \in F_{spq}(M)} E_f ||\hat{f} - f||_{p'}^{p'}.$$

(the infimum being taken over all *estimators (taking their values in a space containing $F_{spq}(M)$). There exists a constant C_3 such that*

$$\begin{array}{rcll} R_n & \ge & C_3(\frac{\log n}{n})^{\alpha p'} & \epsilon \le 0 \\ & \ge & C_3 n^{-\alpha p'} & \epsilon > 0. \end{array}$$

Remarks. 1. As will be shown in the next two sections, the lower bound of Theorem 2 is sharp, at least in the cases ($p' \ge p, 1 < sp < (p'-p)/2$) and ($p = 2, s > 1/p$).

2. We note two special phenomena. First, an "elbow" appears in the rate of convergence: the "usual" rate ($s/(1+2s)$) applies only if s is large enough - in other cases, the rate is $s'/(1+2s-2/p)$. Secondly, a log term appears in the low regularity cases.

3. Comparison with Theorem 1 now shows that linear estimates have sub-optimal rates of convergence for $p' \ge 2,\ p < p',\ s > 1/p$.

PROOF OF THEOREM 2. We give only a brief sketch, as it is a slight modification of Nemirovskii's method to the case of densities. For small s (i.e. $\epsilon \le 0$), we consider the set of vertices of a pyramid

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

where g_0 is some infinitely differentiable density satisfying $g_0 \ge c$ for x in the interval $[-A, A]$ containing the support of ϕ and ψ. Choose M so that $J'_{spq}(g_0) \le M/2$ 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 $k \neq k'$. Finally, in order that $\mathcal{P}_j$ be included in $F_{spq}(M)$, choose γ such that $0 \leq \gamma \leq \Gamma(j; s, p, M)$, where

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

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

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

$$C_j = \{f_\epsilon = g_0 + \sum_{k \in K_j} \gamma \epsilon_k \psi_{jk}, \epsilon_k = \pm 1\}$$

with

$$0 \leq \gamma \leq \frac{C 2^{-j/2}}{||\psi||_\infty} \wedge \frac{M}{2} 2^{-j(s+1/2)}$$

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

PROOF OF THEOREM 1. For the lower bound, we present the details of the proof in the appendix and give on 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 $p' \geq p$, the least favorable points for *linear* estimates over ℓ_p balls are "spikes" – such as the elements of a fixed $\mathcal{P}_j$ as introduced above (compare [DJ, 1990, Section 8 in the Gaussian case). The lower bound is obtained by randomizing over the elements of $\mathcal{P}_j$.

For the upper bound, it suffices to exhibit an estimator attaining the right rate of convergence, for example the "linear wavelet estimator" (c.f. [KP 1992a]):

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

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 that ϕ has regularity $r > s$.

Proposition 1 *([KP 1992a]) For $\pi \geq 1, \sigma < r$, and $f \in F_{\sigma\pi q}(M)$, if $j(n) = [\log_2(n^{\frac{1}{1+2\sigma}})]$, there exists a constant C_4 such that*

$$E_f ||\hat{f}_{n,j(n)} - f||_\pi^\pi \leq C_4 n^{-\frac{\pi\sigma}{1+2\sigma}}.$$

The result is proved in [KP 1992a] for $\pi \geq 2$, but the same argument extends to $\pi \geq 1$ (cf. also (23) below). The upper bound in Theorem 1 is now a consequence of Proposition 1 and the Sobolev embeddings (see Section 2) $B_{spq} \subset B_{s'p'q}$ for $p' \geq p, s - 1/p = s' - 1/p'$, in which we take $\pi = p'$ and $\sigma = s'$.

4 Threshold wavelet estimators

Among non-linear estimators, we propose 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:

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

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

$$TW(x) = \sum_{k \in \mathcal{Z}} \hat{\alpha}_{j_1 k} \phi_{j_1 k}(x) + \sum_{j_1}^{j_0} \sum_{k \in \mathcal{Z}} \tilde{\beta}_{jk} \psi_{jk}(x). \tag{8}$$

Before considering the properties of this estimator, we pause for some motivation. The linear wavelet estimator, LW, (corresponding to $j_0 = j_1$) is not optimal if $p < p'$.This may be explained via the decomposition of the error into bias and stochastic (variance) components. If LW uses level $j(n)$, it has bias of order $2^{-j(n)s'p'}$, 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 to add 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 a supereffi-ciency procedure in the spirit of the Hodges-Lehmann estimator near $\beta_{jk} = 0$. This choice makes sense since the constraint $F_{spq}(M)$ on the function "forces" most of the β_{jk} to be small. We focus on the choice $C(j) = \sqrt{j}$. The first theorem describes the behavior of TW when p, q, s are known. An adaptivity result for unknown p, q, s appears in Section 6.

As before, let $\epsilon = sp - (p' - p)/2$. 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 $p' \ge p \vee 1, s - 1/p > 0$. Suppose that*

$$F_{spq}(M, T) = \{f \in F_{spq}(M) : suppf \subset [-T, +T]\}$$

If $C(j) = \sqrt{j}$, there exist constants $C_5 = C_5(s, p, q, M)$ and K_0 such that if

$$\begin{array}{rcl} 2^{j_1(n)} & \simeq & (n(\log n)^{\frac{p'-p}{p} I\{\epsilon \ge 0\}})^{1-2\alpha} \\ 2^{j_0(n)} & \simeq & (n(\log n)^{-I\{\epsilon \le 0\}})^{\alpha/s'} \end{array} \tag{9}$$

and $K \ge K_0$, then

$$\sup_{f \in F_{spq}(M,T)} E_f ||TW - f||_{p'}^{p'} \le \begin{cases} C_5 (\log n)^{(1-\epsilon/sp)\alpha p'} n^{-\alpha p'} & \epsilon > 0 \\ C_5 (\log n)^{(p'/2 - p/q)_+} (\frac{\log n}{n})^{\alpha p'} & \epsilon = 0 \\ C_5 (\log n/n)^{\alpha p'} & \epsilon < 0. \end{cases} \tag{10}$$

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

Remarks. In the case $\epsilon < 0$, the rate is sharp: the bounds in Theorems 2 and 3 agree. In the other cases, the power of n is correct, but an extra logarithmic term appears (as it does also in the work of Nemirovskii).

The logarithmic term does not appear in the case $p' = 2$ studied by Donoho-Johnstone, and we show in the next section that we can modify $C(j)$ so as to obtain the analog of their result when $p' = 2$. The modification has two disadvantages: firstly $C(j)$ is implicitly specified and is hard to calculate, and second, it depends strongly on (p, s, q, p'). Thus it

will not be of use in the final section, where we construct adaptive procedures. However, adaptive rate optimal procedures can be constructed in the Gaussian case using Stein's unbiased estimate of risk to choose thresholds (Donoho and Johnstone, 1993) and it is natural to conjecture that the argument could be extended to the density case also.

It is also of interest to look at the exponent of this extra log term. In case of $\epsilon > 0$, it is strictly better than $\alpha p'$ and is independent of q, but if $\epsilon = 0$, we see that we have an extra term in addition when q sufficiently large. It turns out (Donoho, Johnstone, Kerkyacharian, Picard, 1993) 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, s, p' and blows up for $\epsilon \to 0$ or $q \to \frac{2p}{p'}$, which accounts for the discontinous nature of the results as presented here.

The number of levels used is proportional to $\log_2 n$: indeed $j_1(n) \sim (1-2\alpha)\log_2 n$ and $j_0(n) \sim (\alpha/s')\log_2 n$. In particular, we note that $j_1(n) < j_0(n)$ unless $p' = p, \epsilon > 0$, 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 condition of compact support is not necessary. It is easy to show that it can be replaced by a domination condition of the following type:

(C) $\exists r > 1$, $\exists \omega : \mathcal{R} \to \mathcal{R}$ symmetric, non-negative, decreasing on $\mathcal{R}^+$, $||\omega||_{1/r} < \infty$ such that $\exists a$ for which $f(x) \le \omega(x-a),\ \forall x \in \mathcal{R}$.

Nevertheless, we do not know if the result is still true without any further condition at all on $F_{spq}(M)$.

PROOF OF THEOREM 5. PRELIMINARIES.

Moment bounds. We recall the following result of Bretagnolle-Huber [BH]: Let $Y_1, \ldots, Y_n$ be i.i.d. random variables with $EY_i = 0, EY_i^2 \le \sigma^2, |Y_i| \le A$. Then there exists c_m such that:

$$\begin{aligned} \text{if } m \ge 2 \quad &, \quad E|n^{-1}\sum Y_i|^m \le c_m\left(\frac{\sigma^2 A^{m-2}}{n^{m-1}} + \frac{\sigma^m}{n^{m/2}}\right) \\ \text{if } 1 \le m \le 2 \quad &, \quad E|n^{-1}\sum Y_i|^m \le \sigma^m n^{-m/2}. \end{aligned} \tag{11}$$

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 $g \in L_2(\mathcal{R})$ be bounded with $\int g^2 = 1$. Define $g_{jk}(x) = 2^{j/2} g(2^j x - k)$,

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

Now apply the Bretagnolle-Huber inequalities to $Y_i = g_{jk}(X_i)$ and set

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

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

$$E|\hat{\gamma}_{jk} - \gamma_{jk}|^m \le c_m\{||f||_\infty^{m/2} + ||f||_\infty (2||g||_\infty)^{m-2}\} n^{-m/2} \tag{12}$$

for all j if $1 \le m \le 2$, and as soon as $n \ge 2^j$ for $m > 2$.

Now it is easy to show that if $f \in F_{spq}(M)$, then

$$||f||_\infty \le (1-2^{-s''q'})^{1/q'} J'_{s''\infty q}(f) \le M(1-2^{-s''q'})^{1/q'}. \tag{13}$$

where $s'' = s - 1/p > 0$ and $1/q + 1/q' = 1$. Consequently, when $f \in F_{spq}(M)$, the bound (12) may be written

$$E|\hat{\gamma}_{jk} - \gamma_{jk}|^m \le c_{bh} n^{-m/2}. \tag{14}$$

where c_{bh} depends as shown at (12) and (13) on $s, p, q, M, ||g||_\infty$ and m.

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

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

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

$$P\{|\hat{\beta}_{jk} - \beta_{jk}| > (K/2)\sqrt{j/n}\} \le 2^{-\gamma j}. \tag{15}$$

For example, the choice $c^2(M,\gamma) = 2||\psi||_\infty M$ suffices if $||\psi||_\infty \ge 1$ and $M \ge 2||\psi||_\infty$.

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 $\pi = p' \wedge 2 \ge 1$

$$\begin{aligned} ||\hat{f}||_{p'}^{p'} &\le C^{p'}(\sum_{j_1}^{j_0} ||D_j\hat{f}||_{p'}^{\pi})^{p'/\pi} & (16)\\ ||D_j\hat{f}||_{p'}^{p'} &\le C^{p'}2^{j(p'/2-1)}\sum_k |\hat{f}_{jk}|^{p'}. & (17)\end{aligned}$$

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

$$S(\gamma) = \sum_{j_1}^{j_0} 2^{j\gamma} \le \begin{cases} c_\gamma 2^{\max(j_0\gamma, j_1\gamma)} & \gamma \ne 0 \\ (j_0 - j_1) & \gamma = 0. \end{cases} \tag{18}$$

From (16), and setting $a = p'/(p'-2)$, we may derive the bound

$$E||\hat{f}||_{p'}^{p'} \le \begin{cases} C^{p'}\sum_{j_1}^{j_0} 2^{j(p'/2-1)}\sum_k E|\hat{f}_{jk}|^{p'} & 1 \le p' \le 2 \\ C^{p'}S(\beta a)^{(p'/2-1)+}\sum_{j_1}^{j_0} 2^{j(p'/2-1-\beta p'/2)}\sum_k E|\hat{f}_{jk}|^{p'} & p' > 2 \end{cases} \tag{19}$$

The first inequality is immediate from (16) and (17). When $p' > 2$, we first apply Hölder's inequality in (16) to obtain

$$(\sum ||D_j \hat{f}||_{p'}^2)^{p'/2} \leq (\sum_{j_1}^{j_0} 2^{j\beta p'/(p'-2)})^{\frac{p'}{2}-1} \sum_{j_1}^{j_0} 2^{-j\beta p'/2} ||D_j \hat{f}||_{p'}^{p'}. \tag{20}$$

Combining (20) with Lemma 1 yields the second inequality in (19). If we adopt the purely formal convention that $S^0 = 1$, then the second inequality in (19) with $\beta = 0$ reduces to the first, and so with this convention, we use (19) for all $p' \geq 1$ below.

COMPLETION OF PROOF. The estimator TW in (8) has two parts: a linear piece $\hat{E}_{j_1(n)}$ and a detail term $\hat{D}_{j_1,j_0}$. Along with a corresponding decomposition of f this yields

$$E_f||TW-f||_{p'}^{p'} \leq 3^{p'-1}(E_f||\hat{E}_{j_1(n)}-E_{j_1(n)}f||_{p'}^{p'}+E_f||\hat{D}_{j_1 j_0}-D_{j_1 j_0}f||_{p'}^{p'}+||f-E_{j_0(n)}f||_{p'}^{p'}). \tag{21}$$

where

$$\begin{aligned} E_j f(x) &= \int \sum_{k\in\mathcal{Z}} \phi_{jk}(y)\phi_{jk}(x)f(y)dy \\ D_{j_1 j_0} f(x) &= \int \sum_{j_1}^{j_0} \sum_{k\in\mathcal{Z}} \psi_{jk}(y)\psi_{jk}(x)f(y)dy. \end{aligned}$$

The third and first terms in (21) are easily estimated. We start with the approximation error. Using the second or fourth characterizations of Besov spaces and the Sobolev embeddings $B_{spq} \subset B_{s'p'q}$, it is easy to see that

$$||f - E_{j_0(n)}f||_{p'}^{p'} \leq C(s,p,q,M)2^{-j_0(n)s'p'}. \tag{22}$$

From the choice of $j_0(n)$, this bound has the rate of convergence specified in (10) if $\epsilon > 0, p' = p$; or $\epsilon = 0, p'/2p \leq 1/q$; or $\epsilon < 0$ and is negligible otherwise.

We turn now to the linear term $E_f||\hat{E}_{j_1(n)} - E_{j_1}f||_{p'}^{p'}$. Using Lemma 1, (14) and the compact support of ϕ, this term is bounded by

$$||\theta_\phi||_{p'}^{p'} 2^{j_1(n)(\frac{p'}{2}-1)} \sum_{k\in\mathcal{Z}} E|\hat{\alpha}_{j_1(n)k} - \alpha_{j_1(n)k}|^{p'} \leq C c_{bh}(T+A)(\frac{2^{j_1(n)}}{n})^{p'/2}, \tag{23}$$

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

To decompose the details term, define

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

We may then write

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

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 (15). We first calculate using this, Hölder's inequality, and (14) that

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

Applying (19) gives

$$E||e_{bs}||_{p'}^{p'} \le C^{p'}\cdot c_{bh} n^{-p'/2}\cdot S(\beta a)^{(p'/2-1)_+} S((1-\beta)p'/2-\gamma/r'). \tag{24}$$

Using the notation of (18), we note that (when $p'>2$)

$$S(\beta a)^r S(b) \le c_{\beta a}^r c_b 2^{(ra+b)j_s}, \quad ra = p'/2$$

where $j_s = j_1(n)$ if $a,b<0$ and $j_s = j_0(n)$ if $a,b>0$. 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 (24) (i.e. both when $p'>2$ and the second when $1\le p'\le 2$) are negative. Thus for $p'\ge 1$,

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

For any choice of $\gamma>0$, this bound is smaller than the linear term in (23) and so is asymptotically negligible.

For the term e_{sb}, apply (19) and (18) 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 deviations bound and the inclusion $B_{s'p'q}\subset B_{s'p'\infty}$,

$$\begin{aligned}\sum_k E|\hat{f}_{jk}|^{p'} &\le \sum_k |\beta_{jk}|^{p'} P(D_{jk}) \le ||\beta_{j\cdot}||_{p'}^{p'} 2^{-\gamma j} \\ &\le C||f||_{s'p'\infty}^{p'} 2^{-j(s'p'+p'/2-1+\gamma)}.\end{aligned} \tag{25}$$

Thus

$$\begin{aligned}E||e_{sb}||_{p'}^{p'} &\le C S(\beta a)^{(\frac{p'}{2}-1)_+} S(-p'(\beta/2+s')-\gamma) M^{p'} \\ &\le C 2^{-j_1(n)(\gamma+s'p')},\end{aligned} \tag{26}$$

after choosing β and γ as described for e_{bs} and exploiting the embedding $B_{s,p,\infty}\subset B_{s',p',\infty}$. This term also is seen to be negligible by taking γ large. For example, the choice $\gamma =$

$\gamma_0 = (\alpha/(1-2\alpha) - s')p'$ makes (26) of exactly the same order as (22). The constant K_0 in Theorem 3 may then be taken as $c(M,\psi)\gamma_0$, specified in (15).

For the term e_{bb}, apply (19) to $\hat{f}_{jk} = |\hat{\beta}_{jk} - \beta_{jk}| I\{k \in \hat{B}_j B_j\}$. In this case, using (14)

$$\begin{aligned} \sum_k E|\hat{f}_{jk}|^{p'} &\leq c_{bh} n^{-p'/2} \sum_{k \in B_j} |\frac{2\beta_{jk}}{K}\sqrt{\frac{n}{j}}|^p \\ &\leq C||\beta_j||_p^p \; j^{-p/2} n^{-(p'-p)/2} \\ &\leq C||f||_{sp\infty}^p 2^{-j(s+1/2-1/p)p} j^{-p/2} n^{-(p'-p)/2}. \end{aligned} \tag{27}$$

In the case $\epsilon \neq 0$, we have, as before from (19) and (18)

$$\begin{aligned} E||e_{bb}||_{p'}^{p'} &\leq \frac{CM^p}{n^{(p'-p)/2}} S(\beta a)^{(\frac{p'}{2}-1)_+} S(-\epsilon - \beta p'/2) \\ &\leq \frac{C}{n^{(p'-p)/2}} \begin{cases} 2^{-j_1 \epsilon} & \epsilon > 0 \\ 2^{-j_0 \epsilon} & \epsilon < 0. \end{cases} \end{aligned}$$

Comparison with the bound (29) below shows that these powers are negligible. In the case $\epsilon = 0$, we have

$$\begin{aligned} E||e_{bb}||_{p'}^{p'} &\leq CM^p \frac{(j_0 - j_1)^{(p'-2)_+/2}}{n^{(p'-p)/2}} \sum_{j_1}^{j_0} j^{-p/2} \\ &\leq CM^p \frac{j_0^{(p' \vee 2 - p)/2}}{n^{(p'-p)/2}}, \end{aligned} \tag{28}$$

since $j_0(n)/j_1(n) \sim p'/(p'-2)$ in the case when $p' > 2$.

Finally, we consider the important and rate determining case e_{ss}, in which $\hat{f}_{jk} = \beta_{jk} I\{k \in \hat{S}_j S_j'\}$. When $p' > 2$, instead of (20), we use (17), and obtain the sharper inequality

$$\begin{aligned} \sum_j ||D_j \hat{f}||_{p'}^2 &\leq C \sum_j 2^{j(1-2/p')} (\sum_{k \in S_j'} |\beta_{jk}|^{p'})^{2/p'} \\ &\leq C(2K)^{(2/p')(p'-p)} \sum_j 2^{j(1-2/p')} (j/n)^{(p'-p)/p'} ||\beta_j||_p^{2p/p'}, \end{aligned}$$

where we have put $|\beta_{jk}|^{p'} = |\beta_{jk}|^{p'-p} |\beta_{jk}|^p$ and noted that $k \in S_j'$ implies $|\beta_{jk}| \leq 2K(j/n)^{\frac{1}{2}}$. Now set $\gamma_j = 2^{j(s+1/2-1/p)} ||\beta_j||_p$: since $f \in F_{spq}(M)$, $||\gamma||_q \leq J'_{spq}(f) \leq M < \infty$. Using (16), we obtain

$$\begin{aligned} E||e_{ss}||_{p'}^{p'} &\leq C(\frac{j_0}{n})^{\frac{p'-p}{2}} (\sum_{j_1}^{j_0} |\gamma_j|^{2p/p'} 2^{-j\epsilon 2/p'})^{p'/2} \\ &\leq C(\frac{j_0}{n})^{\frac{p'-p}{2}} \begin{cases} C \; ||\gamma||_\infty^p \; 2^{\max\{-j_0\epsilon, -j_1\epsilon\}} & \epsilon \neq 0 \\ ||\gamma||_q^p \; j_0^{\; p(\frac{p'}{2p} - \frac{1}{q})_+} & \epsilon = 0. \end{cases} \end{aligned} \tag{29}$$

where for $\epsilon = 0$, we have used monotonicity of $l_p(\mathcal{Z})$ norms and Hölder's inequality, according as $q \le 2p/p'$ or $q > 2p/p'$. That these rates correspond to those announced in Theorem 3 follows from the definitions (9) and the equalities

$$\begin{array}{lll} (p'-p)/2 + \epsilon(1-2\alpha) & = \alpha p' & \text{if } \epsilon \ge 0, \\ (p'-p)/2 + \epsilon\alpha/s' & = \alpha p' & \text{if } \epsilon \le 0. \end{array}$$

When $1 \le p' \le 2$, we have from (19) that

$$\begin{aligned} E||e_{ss}||_{p'}^{p'} &\le C^{p'} \sum_{j_1}^{j_0} 2^{j(p'/2-1)} \sum_{k \in S_j'} |\beta_{jk}|^{p'} \\ &\le C(\frac{j_0}{n})^{(p'-p)/2} \sum_{j_1}^{j_0} 2^{-j\epsilon} |\gamma_j|^p. \end{aligned}$$

where γ_j is as above. Arguing using monotonicity of $\ell_p(\mathcal{Z})$ norms, etc. as above, we obtain

$$E||e_{ss}||_{p'}^{p'} \le C(\frac{j_0}{n})^{(p'-p)/2} \begin{cases} ||\gamma||_\infty^p 2^{\max(-j_0\epsilon, -j_1\epsilon)} & \epsilon \ne 0 \\ ||\gamma||_q^p j_0^{(1-p/q)_+} & \epsilon = 0. \end{cases} \quad \blacksquare$$

5 Quadratic loss and Gaussian approximation

We turn now 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 & Johnstone (1992).

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

$$y_{jk} = \theta_{jk} + \varepsilon z_{jk} \quad j = 0, 1, \cdots; \ k = 0, 1, \cdots, 2^j - 1 \tag{30}$$

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 $\theta \in \Theta_{spq}(M) = \{\theta : ||\theta||_{s,p,q} \le M\}$ where, in this section,

$$||\theta||_{s,p,q}^q = \sum_{j \ge 0} (2^{js'} ||\theta_{j.}||_p)^q \tag{31}$$

and $s' = s + 2^{-1} - p^{-1}$, $||\theta_{j.}||_p^p = \sum_{k=0}^{2^j-1} |\theta_{jk}|^p$. From Donoho and Johnstone (1992), it is known that

$$R_N^*(\sigma n^{-1/2}, \Theta_{s,p,q}(M)) \sim \gamma(\sigma n^{-1/2})(M\sigma^2/n)^{2s/(2s+1)} \tag{32}$$

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

We recall also that co-ordinate-wise threshold estimators can be chosen to be within a bounded factor of being asymptotically minimax. Define a soft threshold rule $\hat{\theta}^\lambda$ by $\{\delta_s(y_{jk}, \ \lambda_j), \ j = 0, 1, \ldots, k = 0, 1, \ldots, 2^j - 1\}$. Then DJ (1991) show that in model (30), there exist absolute constants A_{spq} such that

$$\inf_{\lambda = (\lambda_j)} \sup_{\Theta_{spq}(M)} E_\theta ||\hat{\theta}^\lambda - \theta||^2 \le A_{spq} R_N^*(\varepsilon, \Theta_{spq}(M))(1 + o(1)) \tag{33}$$

Theorem 4 *Suppose that either $p \geq 1$ and $s > p^{-1}$ or that $s = p^{-1}$, $p > 1$. Then there exists $\sigma = \sigma(s,p,q,M)$, $c = c(s,p,T)$ and $C_6 = C_6(s,p,q,M)$ such that*

$$\inf_{\hat{f}_n} \sup_{F_{spq}(M,T)} E_f ||\hat{f}_n - f||_2^2 \leq C_6 R_N^*(\sigma n^{-1/2}, \Theta_{spq}(cM))(1 + o(1)). \tag{34}$$

Estimators of the form (42) 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 $Y_1, \cdots, Y_n$ satisfy $EY_i = 0$, $EY_i^2 = 1, |Y_i| \leq M$ 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 $M^2 n^{-1} \log^3 n \leq c_1$,*

$$E(n^{-1/2} S_n - Z)^2 \leq c_2 M^2 n^{-1}. \tag{35}$$

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,σ^2} denote expectation when $Y \sim N(\beta, \sigma^2)$. If $\sigma < \bar{\sigma}$, then*

$$E_{\beta,\sigma^2}[\delta_s(Y,\lambda) - \beta]^2 \leq 2 E_{\beta,\bar{\sigma}^2}[\delta_s(Y,\lambda) - \beta]^2. \tag{36}$$

To apply the lemmas, fix (j,k) and note that $\hat{\beta}_{jk}$ has mean β_{jk} and variance $n^{-1}\sigma_{jk}^2$, where $\sigma_{jk}^2 = \sigma_{jk}^2(f) = \mathrm{Var}_f\, \psi_{jk}(X)$. 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 $Y_i = (\psi_{jk}(X_i) - \beta_{jk})/\sigma_{jk}$, and note that $|Y_i| \leq 2||\psi||_\infty 2^{j/2}/\sigma_{jk} = M_{jk}$ say. We construct $\hat{\gamma}_{jk} = \beta_{jk} + n^{-1/2}\sigma_{jk} Z_{jk}$ by the following recipe.

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

$$\begin{aligned} T_4 = E[\hat{\beta}_{jk} - \hat{\gamma}_{jk}]^2 &= n^{-1}\sigma_{jk}^2 E[n^{-1/2} S_n - Z_{jk}]^2 & (37) \\ &\leq 4||\psi||_\infty^2 c_2 2^j n^{-2}. & (38) \end{aligned}$$

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

$$T_4 \leq 2\,\mathrm{Var}\,\hat{\beta}_{jk} + 2n^{-1}\sigma_{jk}^2 = 4n^{-1}\sigma_{jk}^2 < 16||\psi||_\infty^2 c_1^{-1} 2^j n^{-2} \log^3 n. \tag{39}$$

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

$$T_4 = E[\hat{\beta}_{jk} - \hat{\gamma}_{jk}]^2 \leq c_4 2^j n^{-2} \log^3 n. \tag{40}$$

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

$$[\delta(\hat{\beta}_{jk}, \lambda) - \beta_{jk}]^2 \leq 2[\delta(\hat{\beta}_{jk}, \lambda) - \delta(\hat{\gamma}_{jk}, \lambda)]^2 + 2[\delta(\hat{\gamma}_{jk}, \lambda) - \beta_{jk}]^2. \tag{41}$$

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

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

where we have used the approximation error bound (40), the variance bound (36), and σ^2 is any common upper bound on σ_{jk}^2. For example, all densities $f \in \mathcal{F}'_{spq}(M)$ are uniformly bounded, say by B_0, and so $\sigma_{jk}^2 \le \int \psi_{jk}^2(x) f(x) dx \le B_0$.

Proof of Theorem 6. It suffices to restrict attention to estimators of the form

$$\hat{f} = \sum_k \hat{\alpha}_{j_1 k} \phi_{j_1 k} + \sum_{j_1}^{j_2} \sum_{k \in \mathcal{Z}} \delta_s(\hat{\beta_{jk}}, \lambda_j) \psi_{jk}. \tag{42}$$

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 maximisation shows that

$$\sup\{T_n(f) \; , \; f \in F_{spq}(M, T)\} = M^2 2^{-2js'}$$

where $s' = s + 1/2 - 1/p$. To bound $S_n(f)$ let $S_j = \{k : |2^{-j}k| < T + A\}$, employ (42) 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,

$$E||\hat{f} - f||_2^2 \le Cn^{-1} + 4 \sum_j \sum_{k \in S_j} r(\delta_{\lambda_j}, \beta_{jk}; \sigma n^{-1/2}) + c_5 2^{2j_2} n^{-2} \log^3 n + M^2 2^{-2js'}. \tag{43}$$

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 (43) is bounded by

$$\sup\{\sum_{j'=j_0}^{\infty} \sum_{|k| \le 2^{j'}} r(\delta_{\bar{\lambda_j}}, \theta_{j'k}; \sigma/\sqrt{n}) \; : \; \theta \in \Theta_{spq}(2^{j_0 s'} M)\},$$

which for appropriate choice of $\bar{\lambda_j}$ is bounded by $A_{spq} R_N^*(\sigma n^{-1/2}, \Theta_{spq}(2^{j_0 s'} M))(1 + o(1))$. Thus for $c = c(s, p, T)$ 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 (43) are of smaller order than R_N^*, namely $n^{-2s/(2s+1)}$ (cf (32)). A sufficient condition for this is easily seen to be

$$\frac{s}{2s+1}\frac{1}{s'}\log_2 n \; << j_2(n) << \frac{s+1}{2s+1}\log_2 n - \frac{3}{2}\log_2\log_2 n \tag{44}$$

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

6 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 s, p, q. Fix an integer r_0 and define a class

$$\mathcal{S} = \{(s,p,q) : (1/p) < s \leq r_0, p \leq p', 1 \leq q \leq \infty\}$$

The modification, denoted ATW, is obtained from compactly supported and $(r_0+1)-$ regular functions ϕ, ψ in (8) simply by specifying $C(j) = \sqrt{j}$ as before, and

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

The constant K is chosen as $c(M,\psi)p'r_0$. Thus, ATW is constructed from TW by maximising over $\mathcal{S}$ the range of levels j over which thresholding occurs in (10).

Theorem 5 *Suppose that $X_1, \ldots, X_n$ are i.i.d with density f of compact support contained in $[-T,T]$, and belonging to some class $F_{spq}(M,T)$, where $(s,p,q) \in \mathcal{S}$. If $p' \geq 1$, then for all $(s,p,q) \in \mathcal{S}$, there exists $C_7(s,p,q,M)$ such that*

$$E_f||ATW - f||_{p'}^{p'} \leq \begin{cases} C_7(\log n/n)^{\alpha p'} & \epsilon \neq 0 \\ C_7(\log n)^{(p'/2-p/q)_+}(\log n/n)^{\alpha p'} & \epsilon = 0. \end{cases} \tag{45}$$

Remark. Although the estimator does not depend on (s,p,q), it is not fully adaptive, since its specification still depends on M and p'. A fully adaptive estimator is possible in the Gaussian white noise case, see ([DJKP]).

PROOF. We modify that of Theorem 3. Consider $f \in F_{spq}(M)$ and define indices $j_i(s,p,q)$ by

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

The index $j_1(s,p,q)$ differs only slightly from that used in Theorem 3, which will be denoted $j_1^*(s,p,q)$.

On $F_{spq}(M)$, the linear and bias terms have rates of convergence no worse than TW:

$$\begin{aligned} E_f||\hat{E}_{j_1} - E_{j_1}f||_{p'}^{p'} &\leq C(\frac{2^{j_1(n)}}{n})^{p'/2} \leq C(\frac{2^{j_1^*(s,p,q)}}{n})^{p'/2} \\ E_f||E_{j_0}f - f||_{p'}^{p'} &\leq C2^{-j_0(n)s'p'} \leq C2^{-j_0(s,p,q)s'p'} \end{aligned}$$

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

For e_{ss}, we use a decomposition,

$$e_{ss} = (\sum_{j_1}^{j_1(spq)} + \sum_{j_1(spq)}^{j_0(spq)} + \sum_{j_0(spq)}^{j_0}) \sum_{k \in \hat{S}_j \cap S_j'} \beta_{jk}\psi_{jk} = e_{ssa} + e_{ssb} + e_{ssc}.$$

The second term is of the form studied in the proof of Theorem 3. When $\epsilon \leq 0$, the value of j_1 plays no role asymptotically, and so the first and second terms may be combined and bounded as in Theorem 3. The third term is bounded as in (22), $||e_{ssc}||_{p'}^{p'} \leq C^{p'}2^{-j_0(spq)s'p'}$.

When $\epsilon > 0$, the value of j_0 plays no role asymptotically, and so the second and third terms may be combined and bounded by (29) as in Theorem 3. The slightly different choice of j_1 made here leads to a bound in terms of $(\log n/n)^{\alpha p'}$. For the first term, we have

$$\begin{aligned} E||e_{ssa}||_{p'} &\leq \sum_{j_1}^{j_1(spq)} 2^{j(1/2-1/p')}\|\beta_{j\cdot}\|_{p'} \leq \sum_{j_1}^{j_1(spq)} 2^{j(1/2-1/p')}2^{j/p'}K\sqrt{\frac{j}{n}} \\ &\leq C\sqrt{\frac{2^{j_1(spq)}j_1(spq)}{n}} \leq C(\frac{\log n}{n})^{\alpha}. \end{aligned}$$

The behavior of the term e_{bb} is a little more delicate. We look first at the case $\epsilon \leq 0$, which, as noted earlier, arises only for $p' > 2$. Applying (19) for $\beta = 0$ with (27) gives

$$\begin{aligned} E||e_{bb}||_{p'}^{p'} &\leq C^{p'}(j_0 - j_1)^{(p'-2)/2}c_{bh}n^{-p'/2}\sum_{j_1}^{j_0} 2^{j(p'/2-1)}\sum_k |\frac{2\beta_{jk}}{K}\sqrt{\frac{n}{j}}|^{p_1} \\ &\leq C^{p'}(\frac{j_0}{n})^{(p'-p_1)/2}\left[\sup_j 2^{j(p'-2)/2p_1}||\beta_j||_{p_1}\right]^{p_1} \end{aligned}$$

We choose $p_1 \in (p,p')$ so that $(p'/p_1) - 1 = s + 1/2 - 1/p$; this choice also yields $(p'-p_1)/2 = \alpha p'$. Since $||\beta_j||_p$ increases as p decreases,

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

When $\epsilon > 0$, we decompose

$$\begin{aligned} e_{bb} &= (\sum_{j_1}^{j_1(spq)} + \sum_{j_1(spq)}^{j_0})\sum_k(\hat{\beta}_{jk} - \beta_{jk})\psi_{jk}I\{k \in \hat{B}_j \cap B_j\} \\ &= e_{bba} + e_{bbb}. \end{aligned}$$

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 (19) along with (23) (applied to $\hat{\beta}_{jk}$ instead of $\hat{\alpha}_{jk}$) to conclude that

$$\begin{aligned} E||e_{bba}||_{p'}^{p'} &\leq C^{p'}S(\beta a)^{(p'/2-1)_+}\sum_{j_1}^{j_1(spq)} 2^{-j\beta p'/2}c_{bh}(T+A)(\frac{2^j}{n})^{p'/2} \\ &\leq C^{p'}(\frac{2^{j_1(spq)}}{n})^{p'/2} \leq C^{p'}n^{-\alpha p'}. \quad \blacksquare \end{aligned}$$

7 Appendix.

7.1 Characterizations of Besov spaces

We list here three further characterisations 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.

1. Minimax viewpoint . Let V be a set of densities incluced in a ball in L_p. We recall the definitions and notations of Section 3 for linear estimators. In particular, let $E^l, l = 1, 2$ be the kernels (4) and (5), and let $E^l_j(f) = \int E^l_j(x,y)f(y)dy$.

Theorem 6 *([KP 1992b]) Let $2 \leq p \leq \infty$, and suppose that V is a set of densities contained in a ball of $L_p(\mathcal{R})$ such that*

(1) There exists $C_2 > 0, s > 0$ such that for all n,

$$\inf_{\hat{f} \in F_n} \sup_{f \in V} E_f ||\hat{f} - f||_p^p \geq C_2 n^{-\frac{sp}{1+2s}}, \tag{46}$$

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 exists a kernel E^1 with k integrable, or E^2 with ϕ localized and sufficiently smooth, and a sequence $j(n)$ such that (for $l = 1$ or 2)

$$\sup_{f \in V} E_f ||\hat{E}^l_{j(n)} - f||_p^p < C n^{-\frac{sp}{1+2s}}, \tag{47}$$

Then V is included in a ball B of $B_{s,p,\infty}$, and the problems have the same complexity: (46) and (47) hold with V replaced by B.

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

2. Approximation theory.

Theorem 7 *([Pe],[KP 1992c]) Let $s > 0, 1 \leq p \leq \infty, 1 \leq q \leq \infty$. Suppose that s is not an integer and set $N = [s]$. Assume that*

(E^1) $\int k(x)|x|^s dx < \infty$, and

$$\int x^j k(x) dx = \delta_{j,0} \qquad j = 0, \ldots, N$$

(E^2) (a) For all u, $|\phi(u)| \leq \Phi(|u|)$ and $\int \Phi(x)|x|^s dx < \infty$. (b) $\phi^{(N+1)}$ exists and satisfies

$$\sum_{k \in \mathcal{Z}} |\phi^{(N+1)}(x-k)| < M \quad \text{for all } x \in \mathcal{R}$$

Then $f \in B_{spq}$ if and only if

$$f \in L_p, \quad \text{and} \quad \epsilon_j = 2^{js}||E_j^l f - f||_p \in l_q(\mathcal{N}). \tag{48}$$

This characterisation in terms of approximation rates is one of the most important properties of Besov spaces. For example, condition (48) is necessary but not sufficient for membership in the classical Sobolev spaces.

3. Modulus of continuity (cf [BL], [M]). Suppose that $0 < s < 1$, $1 \le p, q \le \infty$, and set $\tau_h f(x) = f(x-h)$. Set

$$\begin{aligned} \gamma_{spq}(f) &= \left(\int_{\mathcal{R}} \left(\frac{||\tau_h f - f||_p}{|h|^s} \right)^q \frac{dh}{|h|} \right)^{1/q} \\ \gamma_{sp\infty}(f) &= \sup_{h \in \mathcal{R}} \frac{||\tau_h f - f||_p}{|h|^s}. \end{aligned}$$

In the case $s = 1$, set

$$\begin{aligned} \gamma_{1pq}(f) &= \left(\int_{\mathcal{R}} \left(\frac{||\tau_h f + \tau_{-h} f - 2f||_p}{|h|} \right)^q \frac{dh}{|h|} \right)^{1/q} \\ \gamma_{1p\infty}(f) &= \sup_{h \in \mathcal{R}} \frac{||\tau_h f + \tau_{-h} f - 2f||_p}{|h|}. \end{aligned}$$

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

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

Remarks. (1) It is easy to see from the definitions that $B_{s\infty 1}$ for $s \ge 0$ and $B_{s\infty q}$ for $s > 0, q > 1$ are contained in the space of bounded continuous functions.

2. There are other characterisations of Besov spaces (for example as Lions-Peetre interpolations of Sobolev spaces, or using Littlewood-Paley decompositions, cf [P], [BL]) that we will not need here.

7.2 Lower bound for linear estimators.

We consider a subclass of densities:

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

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 $x \in \mathcal{R}$. 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 $\frac{\partial}{\partial\lambda_{jk}}[E_f\hat{f}_L(x)]$ nor $a_{jk} := \int\frac{\partial}{\partial\lambda_{jk}}[E_f\hat{f}_L(x)]\psi_{jk}(x)dx$ depends on the choice of $f \in \tilde{V}_j$.

We apply an L_1 version of the Cramer-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

$$\begin{aligned} L &= \sum_i\frac{1}{f_\theta(x_i)}\frac{\partial}{\partial\theta}f_\theta(x_i) = \sum_i\frac{\psi_{jk}(x_i)}{f_\theta(x_i)}, \quad \text{and} \\ |L| &\le n||\psi||_\infty/C. \end{aligned}$$

Thus for $p' \ge 1$,

$$\begin{aligned} E_\theta|\hat{T}|^{p'} &\ge (E_\theta|\hat{T}|)^{p'} \\ &\ge C|a_{jk}|^{p'}n^{-p'/2}. \end{aligned} \tag{49}$$

Observe now that if $D_j = E_{j+1}^2 - E_j^2$ (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

$$E_f||\hat{f}_L - f||_{p'}^{p'} \ge a'_{p'}2^{j(\frac{p'}{2}-1)}E_f\sum_k|\hat{\alpha}_{jk} - \lambda_{jk}|^{p'}. \tag{50}$$

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

$$\begin{aligned} R_n^*(\hat{f}_L) &= \sup_{f \in \mathcal{F}_{spq}(M)} E_f ||\hat{f}_L - f||_{p'}^{p'} \geq \frac{1}{\text{card } \mathcal{P}_j} \sum_{f \in \mathcal{P}_j} E_f ||\hat{f}_L - f||_{p'}^{p'} \\ &\geq 2^{-(j+1)} \sum_{k \in K_j} E_{f_k} ||\hat{f}_L - f_k||_{p'}^{p'} + E_{f'_k} ||\hat{f}_L - f'_k||_{p'}^{p'} \\ &\geq a'_{p'} 2^{j(\frac{p'}{2}-1)} 2^{-(j+1)} \sum_{k \in K_j} \{ \sum_{\substack{k' \in K_j \\ k' \neq k}} E_{f_k} |\hat{\alpha}_{jk'}|^{p'} + E_{f'_k} |\hat{\alpha}_{jk'}|^{p'} \\ &\qquad + E_{f_k} |\hat{\alpha}_{jk} - \gamma|^{p'} + E_{f'_k} |\hat{\alpha}_{jk} + \gamma|^{p'} \} \end{aligned} \tag{51}$$

using Lemma 1.

But

$$\begin{aligned} E_{f_k} |\hat{\alpha}_{jk} - \gamma|^{p'} + E_{f'_k} |\hat{\alpha}_{jk} + \gamma|^{p'} &\geq |E_{f_k} \hat{\alpha}_{jk} - \gamma|^{p'} + |E_{f'_k} \hat{\alpha}_{jk} + \gamma|^{p'} \\ &\geq 2^{-(p'-1)} |E_{f_k} \hat{\alpha}_{jk} - E_{f'_k} \hat{\alpha}_{jk} - 2\gamma|^{p'} \\ &= 2\gamma^{p'} |a_{jk} - 1|^{p'}. \end{aligned} \tag{52}$$

using Lemma 4.

Using (49), (51) and (52), we obtain

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

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

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

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

$$(\frac{2^j}{n})^{p'/2} \asymp n^{-\frac{s'p'}{1+2s'}}$$

which establishes the first part of Theorem 1. ∎

7.3 Gaussian approximation for quadratic loss.

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

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

b) for $\beta > 0$, $\quad z \to |\delta_s(\beta - z, \lambda) - \beta|$ 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 $X = \beta + \sigma Z$ for $Z \sim N(0,1)$, and drop explicit reference to λ and s. We now apply these properties in turn:

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

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 σ 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, $\theta \in \mathcal{C}$ satisfies $|\theta| \leq 1$ and may differ at each occurrence. Secondly $c_1, c_2, \cdots$ denote absolute constants.

1^0. 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 $E(n^{-1/2}S_n - Z)^2 \leq 2\alpha^2 + 2c_2M^2(1+\alpha)^2 n^{-1} \leq c_3 M^2 n^{-1}$.

2^0. 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 an $N(0,1)$ variate). We show that Z has the desired approximation by considering in turn large, moderate and small deviations, defined respectively by sets $A_1 = \{w : |w| > \sqrt{a \log n}\}, A_2 = \{1 \leq |w| \leq \sqrt{a \log n}\}$ and $A_3 = \{|w| \leq 1\}$. Indeed, we write

$$\begin{aligned} E(W_n - Z)^2 &= E\{(W_n - Z)^2, |W_n| > \sqrt{a \log n}\} + \int_{A_2 \cup A_3} [w - \Phi^{-1}(F_n(w))]^2 F_n(dw) \\ &= I_1 + I_2 + I_3. \end{aligned} \tag{53}$$

3^0. Small deviations are easily handled by the Berry-Esseen Theorem, which implies that $|r_n(x)| = |F_n(x) - \Phi(x)| \leq C\rho n^{-1/2} \leq CMn^{-1/2}$ since $\rho = E|X_1|^3/(EX_1^2)^{3/2} \leq M$. According to the mean value theorem

$$I_3 \leq \int_{-1}^{1} \frac{r_n^2(w)}{\phi^2(u^\star(w))} F_n(dw) \leq c_3 M^2 n^{-1}, \tag{54}$$

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 $M^2 n^{-1} \log^3 n \leq c_1$.

4^0. For large deviations, first use Hölders inequality to write

$$I_1 \leq c_4(E|W_n|^3 + E|Z|^3)^{2/3} P^{1/3}\{|W_n| > \sqrt{a \log n}\}. \tag{55}$$

Now use Bennett's inequality (see for example Pollard, 1984) to bound

$$\mathrm{P}\left(|S_n| > \sqrt{an \log n}\right) \leq 2 \exp\{-(1/2)a \log n B(M\sqrt{an^{-1} \log n})\} \tag{56}$$

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, $M^2 n^{-1} \log n \leq c_1$, and so the right side is bounded by

$$2 \exp\{-\frac{a}{2} B(\sqrt{c_1 a}) \log n\} \leq 2n^{-3} \tag{57}$$

so long as we choose a large (= 10 say), and c_1 small enough that $aB(\sqrt{c_1 a}) \geq 6$.

Finally, the Bretagnolle - Huber bound (11) shows that

$$E|W_n|^3 \leq c_5(1 + Mn^{-1/2}) \leq c_5(1 + c_1^{1/2}), \tag{58}$$

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

5^0. For moderate deviations, it is sufficient, because of symmetry, to focus on

$$I_2^+ = \int_1^{(a \log n)^{1/2}} [x - \tilde{\Phi}^{-1}(\tilde{F}_n(x))]^2 F_n(dx), \tag{59}$$

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

Lemma 5 *If $x \geq 1$ and $|\tilde{F}/\tilde{\Phi}(x) - 1| \leq e^{-3/2}$, then*

$$|x - \tilde{\Phi}^{-1}(\tilde{F}(x))| \leq x^{-1} e^{3/2} |(\tilde{F}/\tilde{\Phi})(x) - 1| \tag{60}$$

We use also a uniform version of the classical moderate deviations bound based on the Cramer series (cf. Feller (1971), Petrov (1975)). The version we use, due to Sakhanenko (1991), does not require explicit knowledge of the Cramer series $\gamma(x)$. It is phrased instead in terms of the Lyapunov 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, $VarW_n = 1$. Let $x \geq 0$, and $\tilde{F}_n = P(W_n \geq x)$. If*

$$16xL(2x) \leq 1, \tag{61}$$

then the Cramer series $\gamma(x)$ is well-defined and satisfies

$$|\gamma(x)| \leq x^3 L(2x) \tag{62}$$

$$|e^{-\gamma(x)} \tilde{F}_n(x) - \tilde{\Phi}(x)| \leq 32 L(2x) \phi(x). \tag{63}$$

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

$$L(h) \le Mn^{-1/2} e^{hMn^{-1/2}}. \tag{64}$$

The restriction $1 \le x \le \sqrt{a \log n}$ implies that $Mx^3 n^{-1/2}$ and hence $Mxn^{-1/2}$ are both bounded by $(a^3 M^2 n^{-1} \log^3 n)^{1/2} \le 10^{3/2}\sqrt{c_1}$. For a sufficiently small choice of c_1, we may ensure that $|Mx^3 n^{-1/2}| \le 1/18$, say, and hence that condition (61) holds.

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

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

Combining (62) with (64), we conclude that

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

From (63), we obtain

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

The function $\nu(x) = \phi(x)/x\tilde{\Phi}(x)$ is decreasing in $x \ge 0$, and so is bounded below in our case by $\nu(1)$. Combining this with (62) again yields, for $1 \le x \le \sqrt{a \log n}$,

$$\begin{aligned} |\tilde{F}_n/\tilde{\Phi}(x) - 1| &\le c_3(x + x^3)L(2x) \le c_4 x^3 M n^{-1/2} e^{2xMn^{-1/2}} \\ &\le c_5 x^3 M n^{-1/2} \\ &\le c_5 10^{3/2}\sqrt{c_1} \le e^{-3/2}, \end{aligned} \tag{65}$$

again if c_1 is chosen sufficiently small.

Thus Lemma 5 applies also, and from (65)

$$\begin{aligned} I_2^+ &\le e^3 \int_1^{(a \log n)^{1/2}} x^{-2} |(\tilde{F}_n/\tilde{\Phi})(x) - 1|^2 F_n(dx) \\ &\le c_{11} E W_n^4 M^2 n^{-1} \le c_{11}(1 + c_1) M^2 n^{-1} \end{aligned} \tag{66}$$

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

References

[1] Bergh J., Löfström, J. (1976) *Interpolation spaces – An Introduction* Springer. New York.

[2] Bretagnolle, J. and Carol-Huber, C. (1979) Estimation des densités: risque minimax, *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **47** 119-137.

[3] Daubechies, I. (1992) *Ten Lectures on Wavelets* SIAM: Philadelphia.

[4] Devroye, L. (1985) *Nonparametric Density Estimation.* Wiley, New York.

[5] Donoho, D. L. and Johnstone, I. M (1990) Minimax risk over ℓ_p-balls. Technical Report, Department of Statistics, University of California, Berkeley.

[6] Donoho, D. L. and Johnstone, I. M (1992) Minimax Estimation via Wavelet shrinkage. Technical Report, Department of Statistics, Stanford University.

[7] Donoho, D. L. and Johnstone, I. M (1993) Adapting to unknown smoothness via Wavelet shrinkage. Technical Report, Department of Statistics, Stanford University.

[8] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1993) Wavelet Shrinkage: Asymptopia? Manuscript.

[9] Doukhan, P. and Leon, J. (1990) Deviation quadratique d'estimateur de densité par projection orthogonale. Note aux *Comptes Rendus Acad. Sciences Paris* (A) **310** 424-430.

[10] Feller, W. (1971) *An introduction to probability theory and its applications.* , Vol 2. Wiley, New York.

[11] M. Frazier, B. Jawerth, and G. Weiss (1991) *Littlewood-Paley Theory and the study of function spaces.* NSF-CBMS Regional Conf. Ser in Mathematics, **79**. American Math. Soc.: Providence, RI.

[12] Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1992) Estimation d'une densité de probabilité par méthode d'ondelettes. *Comptes Rendus Acad. Sciences Paris* (A) **315** 211-216.

[13] Kerkyacharian, G. and Picard, D. (1992a) Density estimation in Besov Spaces. *Statistics and Probability Letters* **13** 15-24

[14] Kerkyacharian, G. and Picard, D. (1992b) Density Estimation by Kernel and Wavelets methods - optimality of Besov spaces. To appear in *Stat. and Prob. Letters.* Technical Report, Université de Paris VII.

[15] Kerkyacharian, G. and Picard, D. (1992c) Linear Wavelet Methods and other periodic kernel methods. Submitted. Technical Report, Université de Paris VII.

[16] Meyer, Y. (1990a) *Ondelettes.* Paris: Hermann.

[17] Nemirovskii, A.S. (1985) Nonparametric estimation of smooth regression functions. *Izv. Akad. Nauk. SSR Teckhn. Kibernet.* **3**, 50-60 (in Russian). *J. Comput. Syst. Sci.* **23**, 6, 1-11, (1986) (in English).

[18] Nemirovskii, A.S., Polyak, B.T. and Tsybakov, A.B. (1985) Rate of convergence of nonparametric estimates of maximum-likelihood type. *Problems of Information Transmission* **21**, 258-272.

[19] Peetre, J. (1975) *New Thoughts on Besov Spaces.* Duke Univ Math. Ser. **1**.

[20] Petrov, V. V. (1975) *Sums of Independent Random Variables.* Springer, New York.

[21] Pollard D. (1984) *Convergence of stochastic processes.*Springer. New York.

[22] Sakhanenko, A. I. (1991) Berry-Esseen type estimates for Large Deviation Probabilities. *Siberian Mathematical Journal*, **32,** 647 - 656. Translation of *Sibirskii Matematicheskii Zhurnal* , **32 (4),**133-142.

[23] Scott, D. W. (1992) *Multivariate Density Estimation.* Wiley, New York.

[24] Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

[25] Walter, G. G. (1990) Approximation of the Delta function by wavelets. Preprint, U. Wisonsin - Milwaukee.