The Annals of Statistics
1999, Vol. 27, No. 3, 898–924

ADAPTIVE WAVELET ESTIMATION: A BLOCK THRESHOLDING AND ORACLE INEQUALITY APPROACH

BY T. TONY CAI

Purdue University

We study wavelet function estimation via the approach of block thresholding and ideal adaptation with oracle. Oracle inequalities are derived and serve as guides for the selection of smoothing parameters. Based on an oracle inequality and motivated by the data compression and localization properties of wavelets, an adaptive wavelet estimator for nonparametric regression is proposed and the optimality of the procedure is investigated. We show that the estimator achieves simultaneously three objectives: adaptivity, spatial adaptivity and computational efficiency. Specifically, it is proved that the estimator attains the exact optimal rates of convergence over a range of Besov classes and the estimator achieves adaptive local minimax rate for estimating functions at a point. The estimator is easy to implement, at the computational cost of $O(n)$. Simulation shows that the estimator has excellent numerical performance relative to more traditional wavelet estimators.

1. Introduction. Wavelet methods have demonstrated considerable success in nonparametric function estimation in terms of spatial adaptivity, computational efficiency and asymptotic optimality. In contrast to the traditional linear procedures, wavelet methods achieve (near) optimal convergence rates over large function classes such as Besov classes and enjoy excellent mean squared error properties when used to estimate functions that are spatially inhomogeneous. For example, as shown by Donoho and Johnstone (1998), wavelet methods can outperform optimal linear methods, even at the level of convergence rate, over certain Besov classes.

Standard wavelet methods achieve adaptivity through term-by-term thresholding of the empirical wavelet coefficients. There, each individual empirical wavelet coefficient is compared with a predetermined threshold. A wavelet coefficient is retained if its magnitude is above the threshold level and is discarded otherwise. A well-known example of term-by-term thresholding is Donoho and Johnstone's VisuShrink [Donoho and Johnstone (1994)]. VisuShrink is spatially adaptive and the estimator is within a logarithmic factor of the optimal convergence rate over a wide range of Besov classes. VisuShrink achieves a degree of tradeoff between variance and bias contributions to the mean squared error. However, the tradeoff is not optimal. VisuShrink reconstruction is often over-smoothed.

Hall, Kerkyacharian and Picard (1999) considered block thresholding for wavelet function estimation which thresholds empirical wavelet coefficients in

Received March 1998; revised April 1999.

AMS 1991 *subject classifications*. Primary 62G07; secondary 62G20.

Key words and phrases. Adaptivity, Besov space, block thresholding, James–Stein estimator, nonparametric regression, oracle inequality, spatial adaptivity, wavelets.

groups rather than individually. The goal is to increase estimation precision by utilizing information about neighboring wavelet coefficients. The method of Hall, Kerkyacharian and Picard is to first obtain a near unbiased estimate of the sum of squares of the true coefficients within a block and then to keep or kill all the coefficients within the block based on the magnitude of the estimate. The estimator attains the exact minimax rate of convergence without the logarithmic penalty over a range of perturbed Hölder classes [Hall, Kerkyacharian and Picard (1999)].

In the present paper, we study block thresholding rules via the approach of ideal adaptation with oracle. An oracle will not tell us the true estimand, but will tell us, for our method, the "best" choice of smoothing parameters. Ideal adaptation is the performance which can be achieved from smoothing with the aid of an oracle. This approach has been used by Donoho and Johnstone (1994) in developing term-by-term thresholding procedures. Our goal is to derive an estimator that achieves simultaneously three objectives: adaptivity, spatial adaptivity and computational efficiency.

We use the standard device of transforming a function estimation problem into a normal mean problem of estimating the wavelet coefficients in the sequence domain [see, e.g., Donoho and Johnstone (1994)]. After Section 2 in which basic notation and motivations are reviewed, we study in Section 3 the problem of estimating a normal mean by a special family of block shrinkage estimators. The coordinates of the mean vector are estimated in groups and simultaneous shrinkage decisions are made about all coordinates within a block. The performance of the estimators is compared to that of an ideal "estimator" in which case an oracle is available. The goal is to construct estimators which can essentially mimic the performance of an oracle. Among the many traditional shrinkage estimators developed in normal decision theory, the James–Stein estimator is perhaps the best known and is the primary focus in the present paper. A risk inequality for block projection oracle using a blockwise James–Stein rule is derived in Section 3. The block projection oracle inequality offers insights into the balance and tradeoff between block length and threshold level. The oracle inequality, together with the compromise between global and local adaptation, suggests the optimal choice of block size and thresholding constant in wavelet function estimation.

Guided by the oracle inequality developed in Section 3 and motivated by the data compression and the localization properties of wavelets described in Section 2, we define in Section 4 a block thresholding estimator, called *BlockJS*, for nonparametric regression. *BlockJS* overcomes the problem of choosing smoothing parameters and achieves simultaneously the three objectives: adaptivity spatial adaptivity and computational efficiency.

The asymptotic properties of the estimator are investigated in Section 5. The estimator enjoys a high degree of adaptivity and spatial adaptivity in terms of the rate of convergence both for global and local estimation. It is shown that *BlockJS* simultaneously attains the exact optimal rate of convergence over a wide interval of the Besov classes, without prior knowledge of the smoothness of the underlying functions. The estimator automatically adapts to the local smoothness of the underlying function; it attains the local adaptive minimax rate for estimating functions at a point.

BlockJS is easy to implement, at the computational cost of $O(n)$. The estimator is not only quantitatively appealing but visually appealing as well. We show, in Section 5, that *BlockJS* has an interesting smoothness property: if the underlying function is the zero function, then, with probability tending to 1, *BlockJS* is also the zero function.

In Section 6, estimators with different choices of block sizes and threshold levels are compared. Based on the comparisons of the properties of the estimators, it is shown that *BlockJS* achieves the optimal balance between global adaptivity and local adaptivity among the given class of estimators. In the present paper we also suggest that, through the example of the James–Stein estimator, block thresholding serves as a "bridge" between the traditional shrinkage estimators in normal decision theory and the more recent wavelet function estimation. This connection allows us to develop new classes of (near) optimal wavelet estimators, all of which may be useful in different estimation situations.

Simulation results, summarized in Section 7 , show that the estimator has excellent numerical performance relative to more traditional wavelet estimators. For example, for three of the four test functions of Donoho and Johnstone (1994), Doppler, Bumps and Blocks, *BlockJS* has better precision with sample size n than VisuShrink with sample size $2n$ for all n from 512 to 8192 and all signal-to-noise ratios from 3 to 7. Section 8 discusses generalizations and variations of the method. The proofs of the main theoretical results are postponed to Section 9.

2. Notation and motivation.

An orthonormal wavelet basis is generated from dilation and translation of two basic functions, a "father" wavelet ϕ and a "mother" wavelet ψ. In the present paper, the functions ϕ and ψ are assumed to be compactly supported and $\int \phi = 1$. We call a wavelet ψ *r-regular* if ψ has r vanishing moments and r continuous derivatives.

For simplicity in exposition, we work with periodized wavelet bases on $[0, 1]$, letting

$$\phi^p_{jk}(t) = \sum_{l\in\mathscr{Z}} \phi_{jk}(t-l), \qquad \psi^p_{jk}(t) = \sum_{l\in\mathbb{Z}} \phi_{jk}(t-l) \quad \text{for } t \in [0, 1],$$

where

$$\phi_{jk}(t) = 2^{j/2}\phi(2^j t - k), \qquad \psi_{jk}(t) = 2^{j/2}\psi(2^j t - k).$$

The collection $\{\phi^p_{j_0k},\ k = 1, \dots, 2^{j_0};\ \psi^p_{jk},\ j \ge j_0 \ge 0, k = 1, \dots, 2^j\}$ is then an orthonormal basis of $L^2[0, 1]$, provided the primary resolution level j_0 is large enough to ensure that the support of the scaling functions and wavelets at level j_0 is not the whole of $[0, 1]$. The superscript "p" will be suppressed from the notation for convenience.

An orthonormal wavelet basis has an associated exact orthogonal discrete wavelet transform (DWT) that is norm-preserving and transforms sampled data into the wavelet coefficient domain in $O(n)$ steps. We will use the standard device of transforming the problem in the function domain into a problem, in the sequence domain, of estimating the wavelet coefficients. See Daubechies (1992) and Strang (1992) for further details about the wavelets and the discrete wavelet transform.

For a given square-integrable function f on $[0, 1]$, denote $\xi_{jk} = \langle f, \phi_{jk}\rangle$, $\theta_{jk} = \langle f, \psi_{jk}\rangle$. So the function f can be expanded into a wavelet series,

$$f(t) = \sum_{k=1}^{2^{j_0}} \xi_{j_0 k}\phi_{j_0 k}(t) + \sum_{j=j_0}^{\infty}\sum_{k=1}^{2^j} \theta_{jk}\psi_{jk}(t). \tag{2.1}$$

In (2.1), $\xi_{j_0 k}$ are the coefficients at the coarsest level. They represent the gross structure of the function f and θ_{jk} are the wavelet coefficients which represent finer and finer structures of the function f as the resolution level j increases.

A remarkable fact about wavelets is that full wavelet series (those having plenty of nonzero coefficients) represent really pathological functions, whereas normal functions have sparse wavelet series [see Meyer (1992), page 113]. A wavelet transform can compact the energy of a normal function into very few number of large wavelet coefficients. See DeVore, Jawerth and Popov (1992) and Meyer (1992) for details on the data compression property of wavelets.

Wavelet bases are well localized, that is, local regularity properties of a function are determined by its local wavelet coefficients. In particular, a function is smooth at a point if and only if its local wavelet coefficients decay fast enough. The wavelet coefficients at high resolution levels are small where the function is smooth.

Based on these data compression and localization heuristics, one can intuitively envision that all but only a small number of wavelet coefficients of a normal function are negligible and large coefficients at high resolution levels cluster around irregularities of the function.

3. Oracle inequality: a tool.

Suppose we observe a noisy sampled function f,

$$y_i = f(t_i) + \varepsilon z_i, \qquad i = 1, 2, \ldots, n,$$

with $t_i = i/n$, $n = 2^J$ and z_i i.i.d. $N(0, 1)$. We wish to recover the unknown function f based on the sample $y = (y_1, \ldots, y_n)$. By applying the orthogonal discrete wavelet transform to y, we can turn the function estimation problem into a problem of estimating a high-dimensional normal mean. Furthermore, according to the data compression and the localization properties of wavelets, we can envisage the normal mean: the wavelet coefficients, as a sparse vector which contains only a small portion of large coordinates.

We thus consider in this section the problem of estimating a multivariate normal mean. Suppose we are given

$$x_i = \theta_i + \sigma z_i, \tag{3.1}$$

$i = 1, \ldots, n$, $z_i \sim N(0,1)$ i.i.d. with σ known. We wish to estimate $\theta = (\theta_1, \ldots, \theta_n)$ based on the observations $x = (x_1, \ldots, x_n)$ under the mean squared error

$$R(\hat{\theta}, \theta) = \frac{1}{n} \sum E(\hat{\theta}_i - \theta_i)^2. \tag{3.2}$$

When $n = 1$ or 2, decision theory shows that the maximum likelihood estimator x is an admissible estimator of θ. When $n \geq 3$, it was shown by Stein (1955) that x is no longer a good estimator of θ in the sense that x is uniformly dominated by some other estimators. It is known that in order to perform well according to the risk measure (3.2), some form of shrinkage is necessary [see, e.g., Lehmann (1983)].

Diagonal projection oracle. Donoho and Johnstone (1994) studied ideal adaptation using a special class of shrinkage estimators, diagonal projection estimators, in the context of wavelet function estimation. Ideal adaptation is the performance which can be achieved from smoothing with the aid of an oracle. Such an oracle will not tell us the true estimand, but will tell us, for our method, the "best" choice of smoothing parameters [see Donoho and Johnstone (1994)]. The "estimator" obtained with the aid of an oracle is not a true statistical estimator; it represents an ideal for a particular estimation method. The approach of ideal adaptation is to derive true estimators which can essentially "mimic" the performance of an oracle.

Suppose we observe $\{x_i\}$ as in (3.1). Denote by $\mathscr{H}$ a given subset of indices and consider

$$\hat{\theta}_i(\mathscr{H}) = \begin{cases} x_i, & \text{if } j \in \mathscr{H}, \\ 0, & \text{if } j \notin \mathscr{H}. \end{cases}$$

Such a diagonal projection estimator either keeps or omits a coordinate. For each individual coordinate, the expected loss is

$$E(\hat{\theta}_i(\mathscr{H}) - \theta_i)^2 = \sigma^2 I\{i \in \mathscr{H}\} + \theta_i^2 I\{i \notin \mathscr{H}\}.$$

Ideally, to minimize the risk, one would estimate θ_i by x_i when $\theta_i^2 > \sigma^2$ and by 0 otherwise. A diagonal projection (DP) oracle would not tell us the value of θ_i, but would supply exactly the extra side information $\mathscr{H}_{\text{oracle}}(\theta) = \{i: \theta_i^2 > \sigma^2\}$. The ideal diagonal projection consists in estimating only those θ_i larger than the noise level. Supplied with such an oracle, one would have an "estimator" $\hat{\theta}_i^{\text{ideal}} = x_i\ I(\theta_i^2 > \sigma^2)$ and would attain the ideal risk

$$R_{\text{dp.oracle}}(\theta, \sigma, 1) = \frac{1}{n} \sum_{i=1}^{n} \min(\theta_i^2, \sigma^2).$$

The "estimator" $\hat{\theta}_i^{\text{ideal}}$, however, is not a true estimator in a statistical sense, because it depends on the unknown parameter θ. To mimic the performance of the DP oracle, Donoho and Johnstone (1994) proposed the soft threshold estimator,

$$\hat{\theta}_i^* = \operatorname{sgn}(x_i)(|x_i| - \sigma\sqrt{2\log n})_+, \tag{3.3}$$

and showed that the estimator comes (essentially) within a logarithmic factor of the ideal risk for all $\theta \in \mathbb{R}^n$. Specifically, they show the following DP oracle inequality:

$$R(\hat{\theta}^*, \theta) \le (2\log n + 1)[R_{\text{dp.oracle}}(\theta, \sigma, 1) + \sigma^2/n] \quad \text{for all } \theta \in \mathbb{R}^n. \tag{3.4}$$

Donoho and Johnstone derive the DP oracle inequality primarily for wavelet function estimation. They and coauthors show that the wavelet estimator, VisuShrink, achieves unusual adaptivity. The estimator comes within a logarithmic factor of the minimax rates over a wide range of Besov classes [Donoho, Johnstone, Kerkyacharian and Picard (1995)].

Block projection oracle. A DP estimator keeps or kills each coordinate individually without using information about other coordinates. On the contrary, a block projection (BP) estimator thresholds coordinates in groups, it uses information about neighboring coordinates. Simultaneous decisions are made to retain or discard all the coordinates within the same group.

Suppose $\{x_i\}$ are given as in (3.1). Let $B_1, B_2, \dots, B_N$ be a partition of the index set $\{1, \dots, n\}$ with each B_i of size L (For convenience, we assume that the sample size n is divisible by the block size L). Let $\mathscr{H}$ be a subset of the block indices $\{1, \dots, N\}$. A block projection estimator associated with $\mathscr{H}$ is defined as

$$\hat{\theta}_{B_j}(\mathscr{H}) = \begin{cases} x_{B_j}, & \text{if } j \in \mathscr{H}, \\ 0, & \text{if } j \notin \mathscr{H}. \end{cases} \tag{3.5}$$

where x_{B_j} denotes the vector $(x_i)_{i\in B_j}$. For each given block the expected loss is

$$E\|\hat{\theta}_{B_j}(\mathscr{H}) - \theta_{B_j}\|^2 = L\sigma^2 I\{j \in \mathscr{H}\} + \|\theta_{B_j}\|_2^2 I\{j \notin \mathscr{H}\}. \tag{3.6}$$

To minimize the risk (3.6), we would ideally like to choose $\mathscr{H}$ to consist of blocks with signal greater than noise, that is, $\|\theta_{B_j}\|_2^2 > L\sigma^2$. A BP oracle would supply exactly this side information,

$$\mathscr{H}_* = \mathscr{H}_*(\theta) = \{j\colon \|\theta_{B_j}\|_2^2 > L\sigma^2\}.$$

With the aid of the BP oracle, one has the ideal block projection "estimator,"

$$\hat{\theta}_{B_j}(\mathscr{H}_*) = \begin{cases} x_{B_j}, & \text{if } j \in \mathscr{H}_*, \\ 0, & \text{if } j \notin \mathscr{H}_* \end{cases} \tag{3.7}$$

with the ideal risk

$$R_{\text{bp.oracle}}(\theta, \sigma, L) = \inf_{\mathscr{H}} \frac{1}{n} E\|\hat{\theta}(\mathscr{H}) - \theta\|^2 = \frac{1}{n}\sum_{j=1}^{N}(\|\theta_{B_j}\|_2^2 \wedge L\sigma^2), \tag{3.8}$$

where $a \wedge b = \min(a, b)$. It is clear that the ideal risk is unattainable in general because it requires the knowledge of an oracle which is unavailable in most realistic situations. The ideal "estimator" (3.7) is not a true estimator in a statistical sense. Our first goal is to derive a practical estimator which can mimic the performance of the BP oracle. That is, we wish to construct an estimator whose risk is close to the risk of the ideal "estimator." In the present paper, we focus on the well-known James–Stein estimator. Generalizations of the method are discussed in Section 8.

Suppose x_i are observed as in (3.1). James and Stein (1961) proposed a particularly simple shrinkage estimator, $\hat{\theta}^{(1)} = (1-(n-2)\sigma^2/S^2)\,x$ where $S^2 = \sum x_i^2$. James and Stein (1961) show that the estimate dominates the maximum likelihood estimator x when $n \geq 3$. It is easy to see that the estimator $\hat{\theta}^1$ is further dominated by $\hat{\theta}^{(2)} = (1-(n-2)\sigma^2/S^2)_+\,x$.

Efron and Morris (1973) showed that the (positive part) James–Stein estimator $\hat{\theta}^{(2)}$ does more than just demonstrate the inadequacy of the maximum likelihood estimator x. It is a member of a class of good shrinkage rules, all of which may be useful in different estimation problems. The class of estimators, $\hat{\theta} = (1-c\sigma^2/S^2)_+\,x$ can be regarded as truncated empirical Bayes rules. See Efron and Morris (1973).

Under the context of BP estimators, we consider a class of blockwise James–Stein estimators. Within each block B_j, a James–Stein shrinkage rule is applied,

$$\hat{\theta}_{B_j}(L, \lambda) = \left(1 - \frac{\lambda L \sigma^2}{S_j^2}\right)_+ x_{B_j}, \tag{3.9}$$

where $S_j^2 = \|x_{B_j}\|_2^2$. We compare the risk of the estimator (3.9) with the ideal risk (3.8). When the block size L and the threshold λ are properly chosen, the blockwise James–Stein rule can mimic the performance of a BP oracle. Other types of shrinkage procedures are also usable and are presently under consideration; see Berger (1985).

THEOREM 1 (BP oracle inequality). *Assume that x_i and $\hat{\theta}_{B_j}(L, \lambda)$ are given as in* (3.1) *and* (3.9), *respectively. Then*

$$R(\hat{\theta}(L, \lambda), \theta) \leq \frac{1}{n} \sum_{j=1}^{N} (\|\theta_{B_j}\|^2 \wedge \lambda L \sigma^2) + 4\sigma^2 P(\chi_L^2 > \lambda L). \tag{3.10}$$

Written in "oracular" form, we have

$$R(\hat{\theta}(L, \lambda), \theta) \leq \lambda R_{\text{bp.oracle}}(\theta, \sigma, L) + 4\sigma^2 P(\chi_L^2 > \lambda L). \tag{3.11}$$

In particular, with the choice of the block size $L = \log n$ and the threshold $\lambda = \lambda_* \equiv 4.50524$,

$$R(\hat{\theta}(L, \lambda_*), \theta) \leq \lambda_* R_{\text{bp.oracle}}(\theta, \sigma, L) + \frac{2\sigma^2}{n}. \tag{3.12}$$

Therefore, with block size $L = \log n$ and thresholding constant $\lambda_* = 4.50524$, the estimator comes essentially within a constant factor of 4.50524 of the ideal risk. The oracle inequality (3.12) is the main motivation for proposing the *BlockJS* estimator in wavelet function estimation setting. The risk inequality is also a key in proving the asymptotic optimality of the *BlockJS* estimator. The second term on the right-hand side of (3.11) is also important; it determines the balance between the block length L and the threshold level λ when it is applied to function estimation problems. See discussions in Section 6.

REMARK. The threshold $\lambda_* = 4.50524$ is the solution of the equation $\lambda - \log\lambda - 3 = 0$. This particular threshold is chosen so that the corresponding wavelet estimator is (near) optimal in function estimation problems. See Section 6 for further discussions.

Block linear shrinker oracle. An alternative to the block projection estimators given in (3.5) is the more general block linear shrinkers,

$$\hat{\theta}_{B_j} = \gamma_j x_{B_j}, \qquad \gamma_j \in [0, 1].$$

In the case of block projection, $\gamma_j \in \{0, 1\}$. An oracle for block linear shrinkage provides the ideal shrinkage factors $\gamma_j = \|\theta_{B_j}\|_2^2/(\|\theta_{B_j}\|_2^2 + L\sigma^2)$, and it is easy to see that the risk of the ideal "estimator" is given by

$$R_{\text{bls.oracle}} = \frac{1}{n}\sum_{j=1}^{N} \frac{\|\theta_{B_j}\|_2^2 L\sigma^2}{\|\theta_{B_j}\|_2^2 + L\sigma^2}.$$

A risk inequality for the block linear shrinker oracle can be derived as well.

THEOREM 2. *The estimator given in* (3.9) *satisfies*

$$R(\hat{\theta}(L, \lambda), \theta) \le 2\lambda R_{\text{bls.oracle}}(\theta, \sigma, L) + 4\sigma^2 P(\chi_L^2 > \lambda L). \tag{3.13}$$

With $L = \log n$ *and* $\lambda = \lambda_* \equiv 4.505, \ldots,$ *we have*

$$R(\hat{\theta}(L, \lambda_*), \theta) \le 2\lambda_* R_{\text{bls.oracle}}(\theta, \sigma, L) + \frac{2\sigma^2}{n}.$$

Therefore, with block size $L = \log n$ and thresholding constant $\lambda_* = 4.50524$, the risk of the estimator is within a constant factor of the risk of an ideal block linear shrinker.

4. Wavelet shrinkage via the BP oracle inequality: the connection.

Now let us consider the function estimation problem and imagine that the normal mean vector θ in (3.1) is the wavelet coefficients of a regression function. According to the data compression and the localization properties of wavelets, it is reasonable to think of θ as a high-dimensional sparse normal mean vector. We estimate the coordinates of the mean vector in groups by putting the empirical wavelet coefficients into blocks and make simultaneous shrinkage decisions about all coefficients within a block. The true wavelet coefficients at each resolution level is estimated via a blockwise James–Stein rule. The BP oracle inequality (3.12) serves as a guide for choosing the block size and the threshold level, as well as a key in showing the adaptivity of the estimator. We are now ready to define our block thresholding wavelet estimator.

METHOD. Suppose we observe a noisy sampled function f,

$$y_i = f(t_i) + \varepsilon z_i, \qquad i = 1, 2, \dots, n \tag{4.1}$$

with $t_i = i/n$, $n = 2^J$ and z_i i.i.d. $N(0,1)$. The noise level ε is assumed to be known. We are interested in recovering the unknown function f. The precision of an estimator is measured both globally by the expected integrated squared error,

$$R(\hat{f},\ f) = E\|\hat{f} - f\|_2^2, \tag{4.2}$$

and locally by the expected loss at a point,

$$R(\hat{f}(t_0), f(t_0)) = E(\hat{f}(t_0) - f(t_0))^2.$$

Suppose we observe the data $Y = \{y_i\}$ as in (4.1). Let $\tilde{\Theta} = Wn^{-1/2}Y$ be the discrete wavelet transform of $n^{-1/2}Y$. Write

$$\tilde{\Theta} = (\tilde{\xi}_{j_0 1}, \dots, \tilde{\xi}_{j_0 2^{j_0}}, \tilde{\theta}_{j_0 1}, \dots, \tilde{\theta}_{j_0 2^{j_0}}, \dots, \tilde{\theta}_{J-1,1}, \dots, \tilde{\theta}_{J-1,2^{J-1}})^T.$$

Here $\tilde{\xi}_{j_0 k}$ are the gross structure terms at the lowest resolution level, and the coefficients $\tilde{\theta}_{jk}$ ($j = 1, \dots, J-1$, $k = 1, \dots, 2^j$) are fine structure wavelet terms. One may write

$$\tilde{\theta}_{jk} = \theta'_{jk} + n^{-1/2}\varepsilon z_{jk}, \tag{4.3}$$

where the mean θ'_{jk} is approximately the true wavelet coefficients of f, and z_{jk}'s are the transform of the z_i's and so are i.i.d. $N(0,1)$.

At each resolution level j, the empirical wavelet coefficients $\tilde{\theta}_{jk}$ are grouped into nonoverlapping blocks of length L. Let (jb) denote the set of indices of the coefficients in the bth block at level j, that is,

$$(jb) = \{(j,k)\colon (b-1)L + 1 \le k \le bL\}.$$

Let $S_{(jb)} \equiv \sum_b \tilde{\theta}_{jk}^2$ denote the L_2 energy of the noisy signal in block (jb). We then apply the James–Stein shrinkage rule to each block (jb). For $jk \in (jb)$,

$$\hat{\theta}_{jk} = (1 - \lambda L\varepsilon^2/S_{(jb)}^2)_+\, \tilde{\theta}_{jk}. \tag{4.4}$$

Applying the inverse discrete wavelet transform (IDWT), we obtain the estimate of f at the sample points. That is, $\{f(t_i)\colon i = 1, \dots, n\}$ is estimated by $\hat{f} = \{\widehat{f(t_i)}\colon i = 1, \dots, n\}$ with $\hat{f} = W^{-1}n^{1/2}\hat{\Theta}$. The estimate of the whole function f is given by

$$\hat{f}_n(t) = \sum_{k=1}^{2^{j_0}} \tilde{\xi}_{j_0 k}\phi_{j_0 k}(t) + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j} \hat{\theta}_{jk}\psi_{jk}(t). \tag{4.5}$$

Based on the BP oracle inequality we derived in Section 3, we choose the block size $L = \log n$ and the threshold $\lambda = \lambda_* = 4.50524$. With these particular choices of block size and threshold level in (4.4), we call the estimator in (4.5) *BlockJS* and denote the estimator in (4.5) by $\hat{f}_n^*$.

REMARK. The block length $L = \log n$ is selected based on the optimal compromise of global and local adaptivity. The thresholding constant $\lambda_* = 4.50524$ is chosen according to the BP oracle inequality and a minimax criterion discussed in Section 6. With the given block length and threshold level, the estimator achieves both global and local adaptivity simultaneously. See Section 6 for further details on the choices of block length and threshold level.

5. Properties of the BlockJS estimator.

Global properties. As is traditional in the wavelet literature, we investigate the adaptivity of the *BlockJS* procedure across a range of Besov classes. Besov spaces are a very rich class of function spaces. They include many traditional smoothness spaces such as Hölder and Sobolev spaces, as well as function classes of significant spatial inhomogeneity such as the bump algebra and the bounded variation classes. We show that *BlockJS* enjoys excellent adaptivity across a wide range of Besov classes. Full details of Besov spaces are given, for example, in DeVore and Popov (1988).

For a given r-regular mother wavelet ψ with $r > \alpha$, define the sequence seminorm of the wavelet coefficients of a function f by

$$|\theta|_{b^s_{p,q}} = \left(\sum_{j=j_0}^{\infty}\left(2^{js}\left(\sum_k |\theta_{jk}|^p\right)^{1/p}\right)^q\right)^{1/q} \tag{5.1}$$

where $s = \alpha + 1/2 - 1/p$. The wavelet basis provides smoothness characterization of the Besov spaces. It is an important fact that the Besov function norm is equivalent to the sequence norm of the wavelet coefficients of f. See Meyer (1992). We will always use the equivalent sequence norm in our calculations with $\|f\|_{B^\alpha_{p,q}}$. The Besov class $B^\alpha_{p,q}(M)$ is defined to be the set of all functions whose Besov norm is less than M.

Denote the minimax risk over a function class $\mathscr{F}$ by

$$R(\mathscr{F}, n) = \inf_{\hat{f}_n}\sup_{\mathscr{F}} E\|\hat{f}_n - f\|_2^2.$$

Donoho and Johnstone (1998) show that the minimax risk over a Besov class $B^{\alpha}_{p,q}(M)$ is of the order n^{-r} with $r = 2\alpha/(1+2\alpha)$, that is,

$$R(B^{\alpha}_{p,q}(M), n) \asymp n^{-2\alpha/(1+2\alpha)}, \qquad n \to \infty.$$

And the minimax linear rate of convergence is $n^{-r'}$ as $n \to \infty$ with

$$r' = \frac{\alpha + (1/p_- - 1/p)}{\alpha + 1/2 + (1/p_- - 1/p)} \quad \text{where } p_- = \max(p, 2). \tag{5.2}$$

Therefore, the traditional linear methods such as kernel and orthogonal series estimates are suboptimal for estimation over the Besov bodies with $p < 2$.

The following theorem shows that the simple block thresholding rule attains the exact optimal convergence rate over a wide range of the Besov scales.

THEOREM 3. *Suppose the wavelet ψ is r-regular. Then BlockJS satisfies*

$$\sup_{f \in B^{\alpha}_{p,q}(M)} E\|\hat{f}^*_n - f\|^2 \le C n^{-2\alpha/(1+2\alpha)} \tag{5.3}$$

for all $M \in (0, \infty)$, $\alpha \in (0, r)$, $q \in [1, \infty]$ and $p \in [2, \infty]$.

Thus, *BlockJS*, without knowing the a priori degree or amount of smoothness of the underlying function, attains the true optimal convergence rate that one could achieve by knowing the regularity. That is,

$$\sup_{f \in B^{\alpha}_{p,q}(M)} E\|\hat{f}^*_n - f\|^2 \asymp R(B^{\alpha}_{p,q}(M), n) \quad \text{for } p \ge 2.$$

Over the Besov classes with $p < 2$, we have the following theorem.

THEOREM 4. *Assume that the wavelet ψ is r-regular. Then the BlockJS estimator is simultaneously within a logarithmic factor from being minimax for $p < 2$,*

$$\sup_{f \in B^{\alpha}_{p,q}(M)} E\|\hat{f}^*_n - f\|^2 \le C n^{-2\alpha/(1+2\alpha)} (\log n)^{(2/p-1)/(1+2\alpha)} \tag{5.4}$$

for all $M \in (0, \infty)$, $\alpha \in [1/p, r)$, $q \in [1, \infty]$ and $p \in [1, 2)$.

Therefore, by comparing with the minimax linear rate (5.2), *BlockJS* achieves advantages over the traditional linear methods even at the level of rates.

REMARK. Hall, Karkyacharian and Picard (1999) study global adaptivity of their block thresholding procedure over a family of perturbed Hölder classes. It can be easily shown, using the oracle inequality (3.12), that *BlockJS* also achieves optimal convergence rates over a wide range of perturbed Hölder classes.

Local adaptation. We now consider the property of *BlockJS* for estimating functions at a point. For functions of spatial inhomogeneity, the local smoothness of the functions varies significantly from point to point and global risk measures such as (4.2) cannot wholly reflect the performance of estimators at a point. The local risk measure

$$R(\hat{f}(t_0), f(t_0)) = E(\hat{f}(t_0) - f(t_0))^2 \tag{5.5}$$

is used for spatial adaptivity.

We measure the local smoothness of a function at a point by its local Hölder smoothness index. Let us define the local Hölder class $\Lambda^\alpha(M, t_0, \delta)$ as follows. For a fixed point $t_0 \in (0, 1)$ and $0 < \alpha \le 1$,

$$\Lambda^\alpha(M, t_0, \delta) = \{f\colon |f(t) - f(t_0)| \le M\,|t - t_0|^\alpha \text{ for } t \in (t_0 - \delta, t_0 + \delta)\}.$$

If $\alpha > 1$, then

$$\Lambda^\alpha(M, t_0, \delta) = \{f\colon |f^{(\lfloor\alpha\rfloor)}(t) - f^{(\lfloor\alpha\rfloor)}(t_0)| \le M\,|t - t_0|^{\alpha'} \text{ for } t \in (t_0 - \delta, t_0 + \delta)\}$$

where $\lfloor\alpha\rfloor$ is the largest integer less than α and $\alpha' = \alpha - \lfloor\alpha\rfloor$.

It is a well-known fact that for global estimation, it is possible to achieve complete adaptation for free in terms of the convergence rate across a range of function classes. That is, one can do as well when the degree of smoothness is unknown as one could do if the degree of smoothness is known. However, for estimation at a point, one must pay a price for adaptation. The optimal rate of convergence for estimating $f(t_0)$ over function class $\Lambda^\alpha(M, t_0, \delta)$ with α completely known is $n^{-2\alpha/(1+2\alpha)}$. Lepski (1990) and Brown and Low (1996b) showed that one has to pay a price for adaptation of at least a logarithmic factor even when α is known to be one of two values. It is shown that the best one can do is

$$\left(\frac{\log n}{n}\right)^{2\alpha/(1+2\alpha)},$$

when the smoothness parameter α is unknown. We call $(\log n/n)^{2\alpha/(1+2\alpha)}$ the local adaptive minimax rate over the Hölder class $\Lambda^\alpha(M, t_0, \delta)$.

The following theorem shows that *BlockJS* achieves optimal local adaptation with minimal cost.

THEOREM 5. *Suppose the wavelet ψ is r-regular and ϕ has r vanishing moments with $r \ge \alpha$. Let $t_0 \in (0, 1)$ be fixed. Then the BlockJS estimator $\hat{f}_n^*$ satisfies*

$$\sup_{f \in \Lambda^\alpha(M, t_0, \delta)} E(\hat{f}_n^*(t_0) - f(t_0))^2 \le C\left(\frac{\log n}{n}\right)^{2\alpha/(1+2\alpha)} \tag{5.6}$$

REMARK. The choice of $L = \log n$ is important for achieving the optimal local adaptivity. The result does not hold if $L = (\log n)^{1+\delta}$, $\delta > 0$. See Section 6 for further details.

Denoising property. The *BlockJS* procedure is easy to implement, at the computational cost of $O(n)$. Besides the global and local adaptation properties, *BlockJS* has an interesting denoising property which should offer high visual quality of the reconstruction. If the sample contains purely noise without any signal, then, with probability tending to 1, the underlying function is estimated by the zero function.

THEOREM 6. *If the target function is the zero function* $f \equiv 0$, *then with probability tending to* 1 *BlockJS is also the zero function*; *that is*, *there exist universal constants* P_n *such that*

$$P(\hat{f}_n^* \equiv 0 | f \equiv 0) \geq P_n \to 1 \quad \textit{as } n \to \infty. \tag{5.7}$$

6. Choices of block size and threshold level.

In the problem of estimating a function f based on a sample contaminated with noise,

$$y_i = f(t_i) + \varepsilon z_i, \qquad i = 1, 2, \ldots, n,$$

we have three objectives in mind: adaptivity, spatial adaptivity, and computational efficiency. We indicated in the previous sections that, with block size $L = \log n$ and $\lambda = 4.50524$, *BlockJS* achieves the three objectives simultaneously. In particular, Theorems 3, 5 and 6 hold. Naturally, one would ask, How are the block size and the threshold selected? What is the performance of estimators with other choices of block length and threshold level? To answer these questions, let us first look back at the oracle inequality (3.11).

For the purpose of selecting block size and threshold level, let us regard the right-hand side of the oracle inequality (3.11) as the true risk instead of an upper bound. We then choose the thresholding constant of a given block size by minimizing the risk relative to an ideal risk. A similar approach of minimizing an upper bound on the risk has been used in Wahba (1990), page ix). For a chosen block size L, we compare the risk with the ideal risk, $(\|\theta\|_2^2 \wedge L\sigma^2) + \sigma^2/n$ and select the corresponding threshold according to the minimax quantity

$$\lambda_L = \arg\min_{\lambda} \sup_{\theta} \frac{\lambda(\|\theta\|_2^2 \wedge L\sigma^2) + 4\sigma^2 P(\chi_L^2 > \lambda L)}{(\|\theta\|_2^2 \wedge L\sigma^2) + \sigma^2/n}. \tag{6.1}$$

Based on this criterion, we select the thresholding constant λ_L so that the "risk" of the estimator is minimized, in comparison with the ideal risk. The threshold λ_L increases as the block size L decreases. We consider here three interesting cases of block size, $L = 1$, $\log n$ and $(\log n)^{1+\delta}$. The case $L = 1$ is the standard term-by-term thresholding, and the block size of $L = (\log n)^{1+\delta}$ is used in Hall, Kerkyacharian and Picard (1999). [The thresholding rule in Hall, Kerkyacharian and Picard (1999), however, is different from the blockwise James–Stein rule discussed here.] We first determine the corresponding thresholding constant λ_L.

PROPOSITION 1. *Let the threshold* λ_L *be defined as in* (6.1), *then*:

(i) *With block size* $L = \log n$,

$$\lambda_L \sim 4.50524 \quad \textit{as } n \to \infty. \tag{6.2}$$

(ii) *With* $L = 1$,

$$\lambda_L \sim \sqrt{2\log n} \quad \textit{as } n \to \infty. \tag{6.3}$$

(iii) *With* $L = (\log n)^{1+\delta}$, $\delta > 0$,

$$\lambda_L \sim 1 \quad \textit{as } n \to \infty. \tag{6.4}$$

Our choice of threshold $\lambda_L = 4.50524$ used in *BlockJS* is based on (6.2). The choice of block size $L = \log n$ aims at achieving a high degree of both global and local adaptivity. The proof of Proposition 1 uses Lemma 2 on chi-square tail probabilities in Section 9. How are the performances of the block thresholding estimators with parameters (L, λ_L) for $L = 1$ and $L = (\log n)^{1+\delta}$? We summarize the results in the following theorems.

THEOREM 7. *Let* $L = 1$ *and* $\lambda = \sqrt{2\log n}$, *and denote by* $\hat{f}_n^{(1)}$ *the estimator given by* (4.4) *and* (4.5). *Under the conditions of Theorems* 3 *and* 5, $\hat{f}_n^{(1)}$ *satisfies*

$$\text{(6.5) (i)} \qquad \sup_{f \in B^\alpha_{p,q}(M)} E\|\hat{f}_n^{(1)} - f\|^2 \asymp n^{-2\alpha/(1+2\alpha)} (\log n)^{2\alpha/(1+2\alpha)},$$

$$\text{(6.6) (ii)} \qquad \sup_{f \in \Lambda^\alpha(M, t_0, \delta)} E(\hat{f}_n^{(1)}(t_0) - f(t_0))^2 \le C\left(\frac{\log n}{n}\right)^{2\alpha/(1+2\alpha)},$$

$$\text{(6.7)(iii)} \qquad P(\hat{f}_n^{(1)} \equiv 0 | f \equiv 0) \ge P_n \to 1 \quad \textit{as } n \to \infty.$$

That is, the results of Theorems 5 *and* 6, *but not Theorem* 3, *hold for* $\hat{f}_n^{(1)}$.

Therefore, the estimator $\hat{f}_n^{(1)}$ has the noise-free feature and optimal local adaptivity, but not optimal global adaptivity. The extra logarithmic factor in (6.5) is unavoidable because this is a term-by-term thresholding estimator. The shrinkage function $\eta_\lambda^{JS}(x) = (1 - \lambda/x^2)_+ x$ is bounded between the hard threshold $\eta_\lambda^h(x) = xI(|x| > \lambda)$ and the soft threshold $\eta_\lambda^s(x) = \operatorname{sgn}(x)(|x| - \lambda)_+$. The estimator enjoys essentially the same properties as VisuShrink. This estimator has also been studied by Gao (1998).

What if we choose a larger block size? With $L = (\log n)^{1+\delta}$, we have the following.

THEOREM 8. *Denote by* $\hat{f}_n^{(2)}$ *the estimator given by* (4.4) *and* (4.5), *with block size* $L = (\log n)^{1+\delta}$, $\delta > 0$ *and a fixed thresholding constant* $\lambda > 1$. *Under the* *conditions of Theorems* 3 *and* 5, $\hat{f}_n^{(2)}$ *satisfies*

$$\text{(6.8)}\quad \text{(i)}\qquad \sup_{f\in B^{\alpha}_{p,q}(M)} E\|\hat{f}_n^{(2)} - f\|^2 \asymp n^{-2\alpha/(1+2\alpha)};$$

$$\text{(6.9)}\quad \text{(ii)}\qquad \sup_{f\in \Lambda^{\alpha}(M,t_0,\delta)} E(\hat{f}_n^{(2)}(t_0) - f(t_0))^2 \asymp \left(\frac{\log n}{n}\right)^{2\alpha/(1+2\alpha)} (\log n)^{2\alpha\delta/(1+2\alpha)};$$

$$\text{(6.10)(iii)}\qquad P(\hat{f}_n^{(1)} \equiv 0 | f \equiv 0) \ge P_n \to 1 \quad \text{as } n \to \infty.$$

That is, the results of Theorems 3 *and* 6, *but not Theorem* 5, *hold for* $\hat{f}_n^{(2)}$.

In words, the estimator $\hat{f}_n^{(2)}$ achieves global adaptivity, but not optimal local adaptivity. The extra logarithmic factor in (6.9) is because the estimator is not well localized; the block size $L = (\log n)^{1+\delta}$ is too large to achieve optimal local adaptivity in terms of the rate of convergence at a point. We emphasis here that the optimal block size $L = \log n$ is specifically for the family of blockwise James–Stein estimators and for nonparametric regression models. The optimal choice of block size may differ in other situations.

By comparing the asymptotic properties of the estimators, it is clear that only the *BlockJS* estimator, with block size $L = \log n$ and threshold $\lambda = 4.50524$, achieves both global and local adaptivity simultaneously. The block size of $L = \log n$ achieves the optimal compromise between global and local adaptivity in terms of convergence rates.

We have so far focused on the selection of block length and threshold level based entirely on asymptotics. Empirically, other criteria, of course, can also be used for choosing smoothing parameters. A natural choice is to use the principle of minimizing Stein's unbiased risk estimate [Stein (1981)]. This approach has been used by Donoho and Johnstone (1995) in term-by-term thresholding. Ignoring the higher order approximation error, the method can be described as follows.

The positive part James–Stein estimator (4.4) is weakly differentiable; Stein's formula for unbiased estimate of risk yields that

$$\begin{aligned}\mathrm{SURE}(\tilde{\theta}_{j\cdot}, L, \lambda) \equiv 2^j + \sum_b \frac{\lambda^2 L^2 - 2\lambda L(L-2)}{S^2_{(jb)}} I(S^2_{(jb)} > \lambda L)\\ + (S^2_{(jb)} - 2L) I(S^2_{(jb)} \le \lambda L)\end{aligned}$$

is Stein's unbiased risk estimate at resolution level j. Then we can empirically choose the level-dependent block size L_j and threshold level λ_j to be the minimizer of SURE:

$$(L_j, \lambda_j) = \arg\min_{L,\lambda} \mathrm{SURE}(\tilde{\theta}_{j\cdot}, L, \lambda).$$

The performance of this estimator is currently under study. We will report the results elsewhere in the future.

7. Simulation results. We compare the numerical performance of *BlockJS* with Donoho and Johnstone's VisuShrink and SureShrink as well as Coifman and Donoho's translation-invariant (TI) denoising method. SureShrink selects the threshold at each resolution level by minimizing Stein's unbiased estimate of risk at each resolution level. In the simulation, we use the hybrid method proposed in Donoho and Johnstone (1995). The TI denoising method was introduced by Coifman and Donoho (1995).

Eight test functions representing different level of spatial variability are used. The test functions are normalized so that all of the functions have equal s.d.$(f) = 10$. (Formulas and graphs of the test functions are given in the Appendix). *BlockJS*, VisuShrink, SureShrink and TI denoising are applied to noisy versions of the test functions. Sample sizes from $n = 512$ to $n = 8192$ and signal-to-noise ratios (SNR) from 3 to 7 are considered. To save space, we report here only a brief summary of the simulation results. See Cai (1998) for further details.

Simulation shows that *BlockJS* uniformly outperforms VisuShrink in all examples in terms of the mean squared error. For five of the eight test functions, Doppler, Bumps, Blocks, Spikes and Blip, the estimator has better precisions with sample size n than VisuShrink with sample size $2n$ for all n from 512 to 8192 (see Table 1). *BlockJS* also yields better results than TI denoising in most cases, especially when the underlying function has significant spatial variability. *BlockJS* is comparable to SureShrink in terms of the mean squared error. See Table 1 and Figure 1. The *BlockJS* reconstruction does not contain spurious fine-scale structure that is sometimes contained in SureShrink reconstruction [see Cai (1998)].

Different combinations of wavelets and signal-to-noise ratios yield basically the same results. As an illustration of this, Table 1 gives numerical results for SNR = 7 using Daubechies compactly supported wavelet *Symmlet* 8. Table 1 reports the average squared error over 20 replications with sample sizes ranging from $n = 512$ to $n = 8192$. Figure 1 provides a graphical comparison of the mean squared error of *BlockJS* with those of the other three estimators. In Figure 1, the vertical bars represent the ratios of the MSEs of each estimator to the corresponding MSE of *BlockJS*. The higher the bar the better the relative performance of *BlockJS*.

It would be interesting to compare the numerical performance of *BlockJS* with that of the block thresholding estimator of Hall, Kerkyacharian and Picard (1999). However, their method requires the selection of smoothing parameters, block length and threshold level, and no specific criterion is given for choosing the parameters in finite sample cases. We therefore leave explicit numerical comparison for future work.

8. Discussion. *BlockJS* can be modified by averaging over different block centers. Specifically, for each given $0 \le i \le L-1$, partition the indices at each resolution level j into blocks

$$(jb) = \{(j,k)\colon (b-1)L + i + 1 \le k \le bL + i\}.$$

TABLE 1
Mean squared error from 20 *replications* (*SNR* = 7)

n	***BlockJS***	**Visu**	**Sure**	**TI**	***n***	***BlockJS***	**Visu**	**Sure**	**TI**
Doppler					**HeaviSine**				
512	0.756	1.838	0.984	1.438	512	0.370	0.395	0.361	0.323
1024	0.424	1.188	0.564	0.886	1024	0.217	0.290	0.236	0.223
2048	0.236	0.781	0.352	0.541	2048	0.129	0.204	0.138	0.154
4096	0.121	0.424	0.182	0.292	4096	0.099	0.117	0.080	0.091
8192	0.060	0.259	0.105	0.169	8192	0.059	0.078	0.051	0.062
Bumps					**Blocks**				
512	1.758	5.835	1.187	4.034	512	1.562	3.569	1.335	2.746
1024	0.929	3.610	0.977	2.342	1024	0.949	2.290	0.836	1.847
2048	0.528	2.211	0.547	1.354	2048	0.584	1.615	0.648	1.253
4096	0.391	1.160	0.343	0.712	4096	0.501	0.883	0.367	0.696
8192	0.210	0.707	0.210	0.418	8192	0.290	0.620	0.268	0.461
Spikes					**Blip**				
512	0.274	0.502	0.256	0.268	512	0.258	0.455	0.364	0.369
1024	0.149	0.339	0.114	0.155	1024	0.150	0.335	0.235	0.253
2048	0.106	0.265	0.086	0.110	2048	0.090	0.229	0.132	0.161
4096	0.068	0.191	0.060	0.075	4096	0.069	0.139	0.095	0.096
8192	0.053	0.151	0.046	0.055	8192	0.038	0.085	0.053	0.061
Corner					**Wave**				
512	0.170	0.208	0.187	0.152	512	0.395	1.402	0.277	0.339
1024	0.077	0.114	0.086	0.086	1024	0.178	0.782	0.155	0.203
2048	0.040	0.072	0.045	0.054	2048	0.098	0.461	0.092	0.117
4096	0.036	0.036	0.036	0.035	4096	0.060	0.060	0.060	0.028
8192	0.018	0.018	0.018	0.018	8192	0.044	0.044	0.037	0.014

In the original *BlockJS* estimator, we take $i = 0$. Define $\hat{f}_n^{(i)}$ to be the version of $\hat{f}_n^*$ for a given i and set

$$\hat{f}_n^{**} = \sum_{i=0}^{L-1} \hat{f}_n^{(i)} \Big/ L.$$

The estimator $\hat{f}_n^{**}$ often has superior numerical performance, at the cost of higher computational complexity. This technique was also used in Hall, Peneu, Kerkyacharian and Picard (1997).

The James–Stein estimator has been used in wavelet function estimation by Donoho and Johnstone (1995). The estimator, WaveJS, is constructed by applying the James–Stein estimator resolution-level-wise, so it is not local and does not have the spatial adaptivity enjoyed by the *BlockJS* estimator introduced in the present paper. Indeed, the main purpose of Donoho and Johnstone's introduction of the WaveJS procedure is to show that a linear estimator does not perform well even it is an adaptive and nearly ideal linear estimator [see Donoho and Johnstone (1995)].

We have focused on the James–Stein estimator in the present paper. The block thresholding method can be used with other types of shrinkage esti-

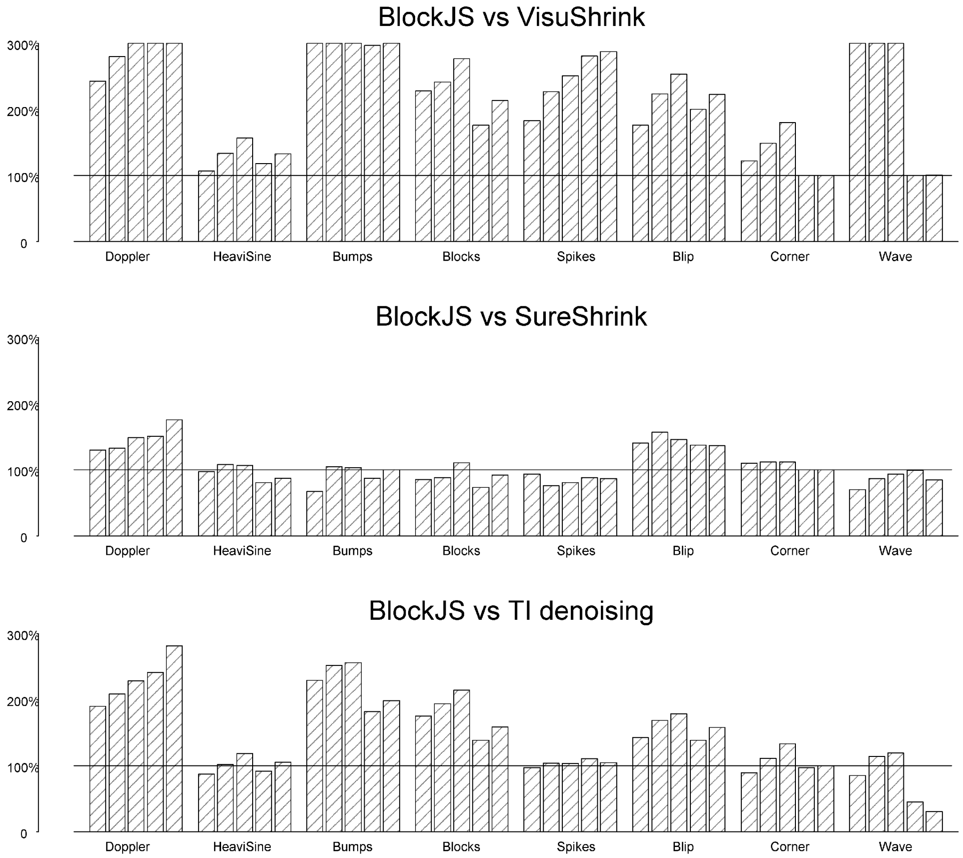

FIG. 1. *Comparison of MSEs. For each signal the bars are ordered from left to right by the sample sizes* ($n = 512$ *to* 8192). *The higher the bar the better the relative performance of BlockJS.*

mators in normal decision theory, for example, estimators of the forms $\hat{\theta} = [1 - \lambda_1\sigma^2/(\lambda_2 + S^2)]_+ y$ or $\hat{\theta} = [1 - c(S^2)/S^2]_+ y$. In this sense, block thresholding serves as a "bridge" between traditional normal decision theory and the recent adaptive wavelet estimation. This bridge enables us to utilize the rich results developed in the decision theory for wavelet function estimation. On the other hand, block thresholding methods can also be used in other statistical function estimation problems such as density estimation and linear inverse problems. We will address these applications elsewhere.

Finally, we note that in the present paper the adaptivity of the *BlockJS* estimator is discussed only in terms of the rate of convergence; the constant factor is not considered. The exact asymptotic risk is unknown for a general Besov class. In the special case of Sobolev classes, Efromovich and Pinsker (1984) constructed an adaptive estimator using Fourier series, which is simultaneously asymptotically sharp minimax, and Efromovich (1998) developed an estimator that is sharp optimal for estimating both the function and its derivatives over Sobolev classes.

9. Proofs.

We shall prove Theorems 1, 3, 4 and 5. Theorem 2 follows from Theorem 1 and some algebra and the proof of Theorem 6 is straightforward. The proof of Theorems 7 and 8 is similar to that of Theorems 3, 4, and 5.

The proof of Theorems 3 and 4 uses the sequence space approach introduced by Donoho and Johnstone (1998). A brief description of the approach and an equivalent result in sequence space is presented in this section. The proof of the equivalent result in a sequence estimation problem is also given in this section.

PROOF OF THEOREM 1. Let $x_i = \mu_i + \sigma z_i$, $i = 1, \dots, L$, and let $\hat{\mu}_i = (1 - \lambda L \sigma^2 / S^2)_+ x_i$, where $S^2 = \|x\|^2$ and $\lambda \geq 1$ is a constant. Denote $R(\hat{\mu}, \mu, \sigma) = E_\sigma \|\hat{\mu} - \mu\|_2^2$, and $\mu^* = \mu/\sigma$. Since $R(\hat{\mu}, \mu, \sigma) = \sigma^2 R(\hat{\mu^*}, \mu^*, 1)$, it suffices to consider only the case $\sigma = 1$ and to show

$$E\|\hat{\theta} - \theta\|_2^2 \leq \|\mu\|^2 \wedge \lambda L + 4P(\chi_L^2 > \lambda L).$$

The (positive part) James–Stein estimator is weakly differentiable; Stein's formula for unbiased estimate of risk yields

$$E\|\hat{\mu} - \mu\|_2^2 = E\,[\mathrm{Sure}(x, L, \lambda)], \tag{9.1}$$

where

$$\begin{aligned} \mathrm{Sure}(x, L, \lambda) = L &+ \frac{\lambda^2 L^2 - 2\lambda L(L-2)}{S^2} I(S^2 > \lambda L) \\ &+ (S^2 - 2L) I(S^2 \leq \lambda L) \end{aligned} \tag{9.2}$$

is Stein's unbiased risk estimate. Simple algebra yields

$$\mathrm{Sure}(x, L, \lambda) \leq \max\{\lambda L - L + 4, L\} \tag{9.3}$$

and it follows from (9.3) trivially,

$$E\|\hat{\mu} - \mu\|_2^2 \leq \lambda L + 4P(\chi_L^2 > \lambda L) \tag{9.4}$$

for $\lambda \geq 1$ and $L \geq 4$. The inequality (9.4) can be verified directly for the cases of $L = 1$, 2 and 3 using the specific noncentral chi-square distributions. For the sake of brevity we omit the proof. It remains to be shown

$$E\|\hat{\mu} - \mu\|_2^2 \leq \|\mu\|^2 + 4P(\chi_L^2 > \lambda L). \tag{9.5}$$

It follows from (9.1) and (9.2) that

$$E\|\hat{\mu} - \mu\|_2^2 = \|\mu\|^2 + E\left[\frac{\lambda^2 L^2 - 2\lambda L^2 + 4\lambda L}{S^2} - S^2 + 2L\right] I(S^2 > \lambda L).$$

Let $\mu_* = \|\mu\|^2/2$, then $E\|\hat{\mu} - \mu\|_2^2$ is a function of μ_*, L and λ. $S^2 = \sum x_i^2$ has a noncentral χ^2-distribution with density

$$f(y) = \sum_{k=0}^{\infty} \frac{\mu_*^k e^{-\mu_*}}{k!} f_{L+2k}(y), \tag{9.6}$$

where $f_m(y) = (1/2^{m/2}\Gamma(m/2))y^{m/2-1}e^{-y/2}$ is the density of a central χ^2-distribution. Write

$$G(\mu_*, L, \lambda) = E\left[\frac{\lambda^2L^2 - 2\lambda L^2 + 4\lambda L}{S^2} - S^2 + 2L\right] I(S^2 > \lambda L).$$

It is easy to see that $G(0, L, \lambda) \le 4P(\chi_L^2 > \lambda L)$. So it suffices to show that $G(\mu_*, L, \lambda)$ is decreasing in μ_*. Denoting by Y_m a central χ^2 variable with degrees of freedom m and using (9.6) as the density of S^2, we have

$$\begin{aligned} G(\mu_*, L, \lambda) &= \sum_{k=0}^{\infty} \frac{\mu_*^k e^{-\mu_*}}{k!} E\left[\frac{\lambda^2L^2 - 2\lambda L^2 + 4\lambda L}{Y_{L+2k}} - Y_{L+2k} + 2L\right] I(Y_{L+2k} > \lambda L) \\ &\equiv \sum_{k=0}^{\infty} \frac{\mu_*^k e^{-\mu_*}}{k!}\, g_k. \end{aligned}$$

Since

$$\frac{\partial G(\mu_*, L, \lambda)}{\partial \mu_*} = \sum_{k=0}^{\infty} \frac{\mu_*^k e^{-\mu_*}}{k!}[g_{k+1} - g_k],$$

it is thus sufficient to show $g_{k+1} - g_k \le 0$ for all $k \ge 0$. Some algebra yields that for $L > 2$,

$$\begin{aligned} g_{k+1} - g_k &= \left(\frac{2\lambda L}{L+2k} - 2\right) P(Y_{L+2k} > \lambda L) \\ &\quad - \frac{2\lambda L(\lambda L - L + 2k + 2)}{(L+2k)(L+2k-2)} P(Y_{L+2k-2} > \lambda L). \end{aligned} \tag{9.7}$$

It is easy to see that $g_{k+1} - g_k \le 0$ when $\lambda \le L + 2k$. For the case of $\lambda > L + 2k$, we appeal to the following lemma on chi-square tail probabilities.

LEMMA 1.

$$P(\chi_{n+2}^2 \ge T) \le \frac{T}{n}\left(1 + \frac{2}{T-n}\right) P(\chi_n^2 \ge T) \quad \textit{if } T > n. \tag{9.8}$$

Applying (9.8) to (9.7), it yields $g_{k+1} - g_k \le 0$ when $\lambda > L + 2k$. Therefore, when $L > 2$, $G(\mu_*, L, \lambda)$ is decreasing in μ_*. Hence $G(\mu_*, L, \lambda) \le G(0, L, \lambda) \le 4P(\chi_L^2 > \lambda L)$, and (9.5) follows:

$$E\|\hat{\mu} - \mu\|_2^2 = \|\mu\|^2 + G(\mu_*, L, \lambda) \le \|\mu\|^2 + 4P(\chi_L^2 > \lambda L).$$

The cases of $L = 1$ and $L = 2$ can be verified directly; we omit the proof here.

Inequality (3.12) follows from (3.11) and the following lemma on the bounds of the tail probability of a central chi-square distribution.

LEMMA 2. *The tail probability of* χ_L^2 *has the following lower and upper bounds*:

$$\tfrac{2}{5}\lambda^{-1} L^{-1/2} [\lambda^{-1} e^{\lambda-1}]^{-L/2} \le P(\chi_L^2 \ge \lambda L) \le \tfrac{1}{2} [\lambda^{-1} e^{\lambda-1}]^{-L/2} \tag{9.9}$$

With $L = \log n$ and $\lambda_* = 4.50524$, Lemma 2 yields

$$P(\chi_L^2 > \lambda L) \le \frac{1}{2}[\lambda^{-1}e^{\lambda-1}]^{-L/2} \le \frac{1}{2n}.$$

The inequality (3.12) now follows from (3.11). □

Asymptotically equivalent estimation problem in sequence space. We shall prove Theorems 3 and 4 by using the sequence space method introduced by Donoho and Johnstone (1998). A key step is to use the asymptotic equivalence results presented by Brown and Low (1996a) and to approximate the problem of estimating f from the noisy observations in (4.1) by the problem of estimating the wavelet coefficient sequence of f contaminated with i.i.d. Gaussian noise.

Donoho and Johnstone (1998) show an equivalence result on the white noise model and the nonparametric regression over the Besov classes $B^{\alpha}_{p,q}(M)$. When the wavelet ψ is r-regular with $r > \alpha$ and $p, q \ge 1$, then a simultaneously near-optimal estimator in the sequence estimation problem can be applied to the empirical wavelet coefficients in the function estimation problem in (4.1), and will be a simultaneously near-optimal estimator in the function estimation problem. For further details about the equivalence and approximation arguments, see Donoho and Johnstone (1995, 1998) and Brown and Low (1996a). For approximation results, see also Chambolle, DeVore, Lee and Lucier (1998).

Under the correspondence between the function and sequence estimation problems, it suffices to consider the following estimation problem in sequence space.

Suppose we observe sequence data,

$$y_{jk} = \theta_{jk} + n^{-1/2}\varepsilon z_{jk}, \qquad j \ge 0,\ k = 1, 2, \dots, 2^j, \tag{9.10}$$

where z_{jk} are i.i.d. $N(0,1)$. The mean vector θ is the object that we wish to estimate. The accuracy of estimation is measured by the expected squared error $R(\hat{\theta}, \theta) = E\sum_{j,k}(\hat{\theta} - \theta)^2$. We assume that θ is known to be in some Besov body $\Theta^s_{p,q}(M) = \{\theta\colon \|\theta\|_{b^s_{p,q}} \le M\}$, where the norm is defined as in (5.1). Make the usual calibration $s = \alpha + 1/2 - 1/p$. The minimax rate for estimating θ over the Besov body $\Theta^s_{p,q}(M)$ is $n^{-2\alpha/(1+2\alpha)}$ as $n \to \infty$ [see Donoho and Johnstone (1998)].

We now apply a *BlockJS*-type procedure to this sequence estimation problem. Let $J = [\log_2 n]$. Divide each resolution level $j_0 \le j < J$ into nonoverlapping blocks of length $L = [\log n]$. Again denote (jb) the bth block at level j. Now estimate θ by $\hat{\theta}^*$ with

$$\hat{\theta}^*_{jk} = \begin{cases} y_{jk}, & \text{for } j \le j_0, \\ (1 - \lambda_* L n^{-1}\varepsilon^2/S^2_{(jb)})_+ \, y_{jk}, & \text{for } jk \in (jb),\ j_0 \le j < J, \\ 0, & \text{for } j \ge J. \end{cases} \tag{9.11}$$

We have the following minimax results for this estimator.

THEOREM 9. *Let $\hat{\theta}^*$ be given as in* (9.11), *then*

$$\sup_{\Theta^s_{p,q}(M)} E\|\hat{\theta}^* - \theta\|_2^2 \tag{9.12}$$
$$\leq \begin{cases} Cn^{-2\alpha/(1+2\alpha)}, & \text{for } p \geq 2, \\ Cn^{-2\alpha/(1+2\alpha)}(\log n)^{(2-p)/(p(1+2\alpha))}, & \text{for } p < 2 \text{ and } \alpha p \geq 1. \end{cases}$$

The results of Theorems 3 and 4 follow from this theorem and the equivalence argument. See also Donoho and Johnstone (1995).

PROOF. We begin by stating the following elementary inequalities without proof.

LEMMA 3. *Let $x \in \mathbb{R}^m$ and $0 < p_1 \leq p_2 \leq \infty$. Then the following inequalities hold*:

$$\|x\|_{p_2} \leq \|x\|_{p_1} \leq m^{1/p_1 - 1/p_2}\|x\|_{p_2}. \tag{9.13}$$

Let y and $\hat{\theta}^*$ be given as in (9.10) and (9.11), respectively. Then,

$$\begin{aligned} E\|\hat{\theta}^* - \theta\|_2^2 &= \sum_{j<j_0}\sum_k E(\hat{\theta}^*_{jk} - \theta_{jk})^2 + \sum_{j=j_0}^{J-1}\sum_k E(\hat{\theta}^*_{jk} - \theta_{jk})^2 + \sum_{j=J}^{\infty}\sum_k \theta^2_{jk} \\ &\equiv S_1 + S_2 + S_3. \end{aligned} \tag{9.14}$$

Denote by C a generic constant that may vary from place to place. It is clear that the first term S_1 is small,

$$S_1 = 2^{j_0}n^{-1}\varepsilon^2 = o(n^{-2\alpha/(1+2\alpha)}). \tag{9.15}$$

First consider the case $p \geq 2$. Since $\theta \in \Theta^\alpha_{p,q}(M)$, so $2^{js}(\sum_{k=1}^{2^j}|\theta_{jk}|^p)^{1/p} \leq M$. Lemma 3 yields that for $p \geq 2$, $\sum_{k=1}^{2^j}|\theta_{jk}|^2 \leq M^2 2^{-j2\alpha}$. Hence,

$$S_3 = \sum_{j=J}^{\infty}\sum_{k=1}^{2^j}\theta^2_{jk} \leq \sum_{j=J}^{\infty} M^2 2^{-j2\alpha} \leq Cn^{-2\alpha} = o(n^{-2\alpha/(1+2\alpha)}). \tag{9.16}$$

Now let us consider the term S_2. Denote by $\beta^2_{(jb)} = \sum_{k\in(jb)}\theta^2_{jk}$ the sum of squared coefficients within the block (jb). Let $J_1 = [(1/(1+2\alpha))\log_2 n]$. So, $2^{J_1} \approx n^{1/(1+2\alpha)}$. The BP oracle inequality (3.12) yields

$$\begin{aligned} S_2 &= \sum_{j=j_0}^{J-1}\sum_k E(\hat{\theta}^*_{jk} - \theta_{jk})^2 \leq \lambda_* \sum_{j=j_0}^{J-1}\sum_b(\beta^2_{(jb)} \wedge Ln^{-1}\varepsilon^2) + 2n^{-1}\varepsilon^2 \\ &\leq \lambda_* \sum_{j=j_0}^{J_1-1}\sum_b Ln^{-1}\varepsilon^2 + \lambda_* \sum_{j=J_1}^{J-1}\sum_b \beta^2_{(jb)} + 2n^{-1}\varepsilon^2 \leq Cn^{-2\alpha/(1+2\alpha)}. \end{aligned} \tag{9.17}$$

By combining (9.17) with (9.15) and (9.16), we have

$$E\|\hat{\theta}^* - \theta\|_2^2 \leq Cn^{-2\alpha/(1+2\alpha)} \quad \text{for } p \geq 2.$$

Now let us consider the case $p < 2$. Since $\theta \in \Theta^{\alpha}_{p,q}(M)$ and $p < 2$, Lemma 3 yields $\sum_{k=1}^{2^j} |\theta_{jk}|^2 \le M^2 2^{-j2s}$. The assumption $\alpha p \ge 1$ implies that S_3 is of higher order,

$$S_3 = \sum_{j=J}^{\infty} \sum_{k=1}^{2^j} \theta_{jk}^2 \le \sum_{j=J}^{\infty} M^2 2^{-j2s} \le Cn^{-2\alpha-1+2/p} = o(n^{-2\alpha/(1+2\alpha)}). \tag{9.18}$$

Now we consider the term S_2. First we state the following lemma without proof.

LEMMA 4. *Let* $0 < p < 1$ *and* $S = \{x \in \mathbb{R}^k\colon \sum_{i=1}^{k} x_i^p \le B,\ x_i \ge 0,\ i = 1, \dots, k\}$. *Then for* $A > 0$,

$$\sup_{x \in S} \sum_{i=1}^{k} (x_i \wedge A) \le BA^{1-p}.$$

Again denote $\beta_{(jb)}^2 = \sum_{k \in (jb)} \theta_{jk}^2$. The BP oracle inequality (3.12) yields

$$S_2 = \sum_{j=j_0}^{J-1} \sum_k E(\hat{\theta}_{jk}^* - \theta_{jk})^2 \le \lambda_* \sum_{j=j_0}^{J-1} \sum_b (\beta_{(jb)}^2 \wedge Ln^{-1}\varepsilon^2) + 2n^{-1}\varepsilon^2. \tag{9.19}$$

Let J_2 be an integer satisfying $2^{J_2} \asymp n^{1/(1+2\alpha)}(\log n)^{(2-p)/p(1+2\alpha)}$. Then

$$\begin{aligned} \lambda_* \sum_{j=j_0}^{J_2-1} \sum_b (\beta_{(jb)}^2 \wedge Ln^{-1}\varepsilon^2) &\le \sum_{j=j_0}^{J_2-1} \sum_b \lambda_* Ln^{-1}\varepsilon^2 \\ &\le Cn^{-2\alpha/(1+2\alpha)}(\log n)^{(2-p)/p(1+2\alpha)}. \end{aligned} \tag{9.20}$$

Note that $\sum_b (\beta_{(jb)}^2)^{p/2} \le \sum_k (\theta_{j,k}^2)^{p/2} \le M 2^{-jsp}$. Lemma 4 yields

$$\lambda_* \sum_{j=J_2}^{J-1} \sum_b (\beta_{(jb)}^2 \wedge Ln^{-1}\varepsilon^2) \le Cn^{-2\alpha/(1+2\alpha)}(\log n)^{(2-p)/p(1+2\alpha)}. \tag{9.21}$$

We complete the proof by putting (9.15) and (9.18)–(9.21) together:

$$E\|\hat{\theta}^* - \theta\|_2^2 \le Cn^{-2\alpha/(1+2\alpha)}(\log n)^{(2-p)/p(1+2\alpha)}. \qquad \square$$

PROOF OF THEOREM 5. For simplicity, we give the proof for Hölder classes $\Lambda^{\alpha}(M)$ instead of local Hölder classes $\Lambda^{\alpha}(M, t_0, \delta)$. Also we will ignore the fact that the mean $\theta'_{j,k}$ in (4.3) is not exactly, but only approximately, the true wavelet coefficients of f. The approximation error is of higher order than the minimax risk, therefore it is negligible. See Cai and Brown (1998) for details.

First note that for Hölder classes $\Lambda^{\alpha}(M)$ there exists a constant $C > 0$ such that for all $f \in \Lambda^{\alpha}(M)$,

$$|\theta_{j,k}| = |\langle f, \psi_{j,k} \rangle| \le C 2^{-j(1/2+\alpha)}. \tag{9.22}$$

The proof of the theorem makes use of the following elementary inequality.

LEMMA 5. *Let X_i be random variables, $i = 1, \dots, n$. Then*

$$E\biggl(\sum_{i=1}^{n} X_i\biggr)^2 \le \biggl(\sum_{i=1}^{n}(EX_i^2)^{1/2}\biggr)^2. \tag{9.23}$$

Now applying inequality (9.23), we have

$$\begin{aligned} &E(\hat{f}_n^*(t_0) - f(t_0))^2 \\ &\quad = E\Biggl[\sum_{k=1}^{2^{j_0}}(\hat{\xi}_{j_0k} - \xi_{j_0k})\phi_{j_0k}(t_0) + \sum_{j=j_0}^{\infty}\sum_{k=1}^{2^j}(\hat{\theta}_{jk} - \theta_{jk})\psi_{jk}(t_0)\Biggr]^2 \\ &\quad \le \Biggl[\sum_{k=1}^{2^{j_0}}(E(\hat{\xi}_{j_0k} - \xi_{j_0k})^2\phi_{j_0k}^2(t_0))^{1/2} \\ &\qquad + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j}(E(\hat{\theta}_{jk} - \theta_{jk})^2\psi_{jk}^2(t_0))^{1/2} + \sum_{j=J}^{\infty}\sum_{k=1}^{2^j}|\theta_{jk}\psi_{jk}(t_0)|\Biggr]^2 \\ &\quad \equiv (Q_1 + Q_2 + Q_3)^2. \end{aligned}$$

Since we are using wavelets of compact support, there are at most N basis functions ψ_{jk} at each resolution level j that are nonvanishing at t_0, where N is the length of the support of the wavelets ϕ and ψ. Denote $K(t_0, j) = \{k\colon \psi_{j,k}(t_0) \ne 0\}$. Then $|K(t_0, j)| \le N$. It is easy to see that both Q_1 and Q_3 are small,

$$Q_1 = \sum_{k=1}^{2^{j_0}}(E(\hat{\xi}_{j_0k} - \xi_{j_0k})^2)^{1/2}|\phi_{j_0k}(t_0)| = O(n^{-1}), \tag{9.24}$$

$$Q_3 = \sum_{j=J}^{\infty}\sum_{k=1}^{2^j}|\theta_{jk}||\psi_{jk}(t_0)| \le \sum_{j=J}^{\infty} N\|\psi\|_\infty 2^{j/2}C2^{-j(1/2+\alpha)} \le Cn^{-\alpha}. \tag{9.25}$$

We now consider the second term Q_2. Applying Lemma 3 and the BP oracle inequality (3.12), and using (9.22), we have

$$\begin{aligned} Q_2 &\le \sum_{j=j_0}^{J-1}\sum_{k\in K(t_0,\, j)} 2^{j/2}\|\psi\|_\infty(E(\hat{\theta}_{jk} - \theta_{jk})^2)^{1/2} \\ &\le C\sum_{j=j_0}^{J-1} 2^{j/2}[(2^{-j(1+2\alpha)} \wedge Ln^{-1}\varepsilon^2) + Ln^{-2}\varepsilon^2]^{1/2} \\ &\le C(\log n/n)^{\alpha/(1+2\alpha)}. \end{aligned} \tag{9.26}$$

Combining (9.24), (9.25) and (9.26), we have

$$E(\hat{f}_n^*(t_0) - f(t_0))^2 \le C(\log n/n)^{2\alpha/(1+2\alpha)}. \qquad \square$$

APPENDIX

Four of the eight test functions, Doppler, HeaviSine, Bumps and Blocks are from Donoho and Johnstone (1994). Blip and Wave are from Marron, Adak, Johnstone, Neumann and Patil (1998). All of the test functions are normalized so that each function has $s.d.(f) = 10$. Formulas of Spikes and Corner are given below.

$$\begin{aligned}\text{Spikes: } f(x) = 15.6676\big[&e^{-500(x-0.23)^2} + 2\,e^{-2000(x-0.33)^2} + 4\,e^{-8000(x-0.47)^2}\\ &+ 3\,e^{-16000(x-0.69)^2} + e^{-32000(x-0.83)^2}\big].\end{aligned}$$

$$\begin{aligned}\text{Corner: } f(x) = 62.387\big[&10x^3(1-4x^2)I_{(0,\,0.5]}(x) + 3(0.125 - x^3)x^4 I_{(0.5,\,0.8]}(x)\\ &+ 59.443(x-1)^3 I_{(0.8,\,1]}(x)\big]\end{aligned}$$

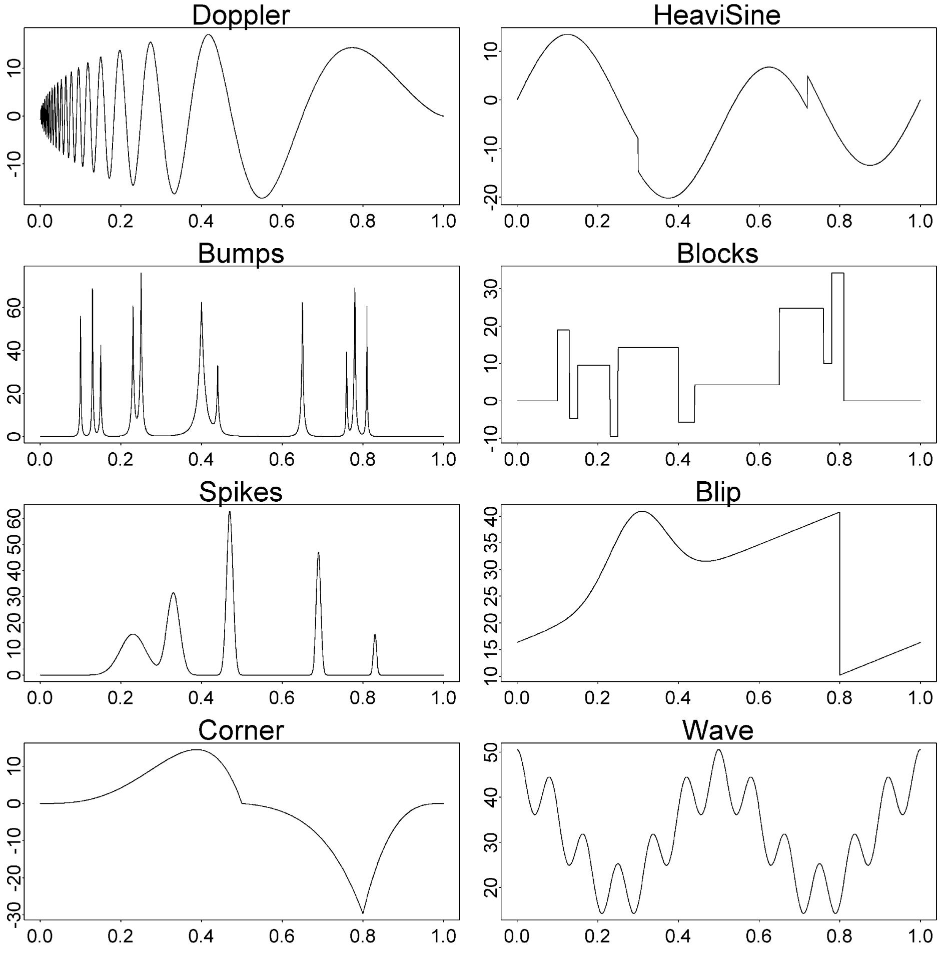

FIG. 2. *Test functions.*

Acknowledgments. It is a pleasure to acknowledge helpful comments by the Editor, James Berger, and Anirban DasGupta which significantly improved the presentation of the paper. The author also thanks two anonymous referees for their thorough and useful comments.

REFERENCES

BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York.

BROWN, L. D. and LOW, M. G. (1996a). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398.

BROWN, L. D. and LOW, M. G. (1996b). A constrained risk inequality with applications to nonparametric functional estimations. *Ann. Statist.* **24** 2524–2535.

CAI, T. (1998). Numerical comparisons of BlockJS estimator with conventional wavelet methods. Unpublished manuscript.

CAI, T. and BROWN, L. D. (1998). Wavelet shrinkage for nonequispaced samples. *Ann. Statist.* **26** 1783–1799.

CHAMBOLLE, A., DEVORE, R., LEE, N. and LUCIER, B. (1998). Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage. *IEEE Trans. Image Processing* **70** 319–335.

COIFMAN, R. R. and DONOHO, D. L. (1995). Translation invariant denoising. *Wavelets and Statistics. Lecture Notes in Statist.* **103** 125–150. Springer, New York.

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

DEVORE, R., JAWERTH, B. and POPOV, V. (1992). Compression of wavelet decompositions. *Amer. J. Math.* **114** 737–785.

DEVORE, R. and POPOV, V. (1988). Interpolation of Besov spaces. *Trans. Amer. Math. Soc.* **305** 397–414.

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **81** 425–455.

DONOHO, D. L. and JOHNSTONE, I. M. (1995). Adapt to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200–1224.

DONOHO, D. L. and JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26** 879–921.

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

EFROMOVICH, S. Y. and PINSKER, M. S. (1984). Self learning algorithm of nonparametric filtration. *Automat. i Telemeh.* **11** 58–65. (In Russian.)

EFROMOVICH, S. Y. (1998). Simultaneous sharp estimation of functions and their derivatives. *Ann. Statist.* **26** 273–278.

EFRON, B. and MORRIS, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130.

GAO, H.-Y. (1998). Wavelet shrinkage denoising using the non-negative garrote. *J. Comput. Graph. Statist.* **7** 469–488.

HALL, P., KERKYACHARIAN, G. and PICARD, D. (1999). On the minimax optimality of block thresholded wavelet estimators. *Statist. Sinica* **9** 33–50.

HALL, P., PENEV, S., KERKYACHARIAN, G. and PICARD, D. (1997). Numerical performance of block thresholded wavelet estimators. *Statist. Comput.* **7** 115–124.

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 361–380. Univ. California Press, Berkeley.

LEHMANN, E. L. (1983). *Theory of Point Estimation.* Wiley, New York.

LEPSKI, O. V. (1990). On a problem of adaptive estimation on white Gaussian noise. *Theory Probab. Appl.* **35** 454–466.

MARRON, J. S., ADAK, S., JOHNSTONE, I. M., NEUMANN, M. H. and PATIL, P. (1998). Exact risk analysis of wavelet regression. *J. Comput. Graph. Statist.* **7** 278–309.

MEYER, Y. (1992). *Wavelets and Operators*. Cambridge Univ. Press.

STEIN, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Stat. Probab.* **1** 197–206. Univ. California Press, Berkeley.

STEIN, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151.

STRANG, G. (1992). Wavelet and dilation equations: a brief introduction. *SIAM Rev.* **31** 614–627.

WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47907
E-MAIL: tcai@stat.purdue.edu