The Annals of Statistics
2003, Vol. 31, No. 1, 110–151

WAVELET THRESHOLDING FOR NON-NECESSARILY GAUSSIAN NOISE: IDEALISM

BY R. AVERKAMP[1] AND C. HOUDRÉ[2]

Freiburg University and Université Paris XII and Georgia Institute of Technology

For various types of noise (exponential, normal mixture, compactly supported, ...) wavelet thresholding methods are studied. Problems linked to the existence of optimal thresholds are tackled, and minimaxity properties of the methods also analyzed. A coefficient dependent method for choosing thresholds is also briefly presented.

1. Introduction. A common underlying assumption in nonparametric curve/surface/signal estimation is that the function to estimate has some redundancy and this is often reflected by the hypothesis that it belongs to a particular functional class. A similar prior assumption is that limited information is present in this curve/surface/signal. For example, it could be discontinuous but only at a limited number of places, or the function to estimate is assumed to have only one mode or to be monotone. Then, the heuristic for the use of wavelets in non-parametric estimation is that the expansion of such a function in a wavelet basis is sparse, that is, only a few of the wavelet coefficients are big and the rest are small and thus negligible. Hence, in order to estimate the function, one has to estimate the large wavelet coefficients and discard the rest. This approach has proved useful and successful as shown, in recent years, by various authors [1, 7, 9–13, 15, 21, 23–25, 28, 33, 39].

Since we do not review the theory of wavelets here, we refer the reader to the books of Daubechies [8] and Meyer [31, 32] for an introduction to the subject. Nevertheless, let us just say that in using a multiresolution approach there is a large family of wavelets with compact support generating orthonormal bases. Moreover, properties of compactly supported wavelets are at the root of a very efficient analog of the fast Fourier transform, the so-called fast wavelet transform. *With this in mind and from now on, we use an orthonormal wavelet basis from a multiresolution analysis adapted to an interval.*

Next, nonparametric estimation via wavelet methods is usually divided into two steps. The first step transforms the data into something which can be input into the fast wavelet transform, that is, noisy versions (denoted by $\widetilde{c}_{j_0,k}$) of the

Received December 1999; revised January 2002.

[1]Supported in part by NSF Grant DMS-96-32032 while this author was visiting the Southeast Applied Analysis Center, School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332.

[2]Supported in part by NSF Grant DMS-98-03239.

AMS 2000 subject classifications. Primary 62G07, 62C20; secondary 60G70, 41A25.

Key words and phrases. Wavelets, thresholding, minimax.

scaling coefficients $c_{j_0,k}$, with j_0 large. The fast wavelet transform is then applied to this data giving noisy versions of the wavelet coefficients $d_{j,k}$ (denoted by $\widetilde{d}_{j,k}$ and called the empirical wavelet coefficients). In the second step, estimates $\widehat{d_{j,k}}$ of the $d_{j,k}$ are computed using the $\widetilde{d}_{j,k}$ and using the heuristic that the wavelet transform of the signal is sparse and that the noise is evenly spread over the empirical wavelet coefficients. From this, estimation of the original function can easily be obtained.

A simple approach, which can be viewed as a first-order approximation, for getting the $\widetilde{c}_{j_0,k}$ is to assume that they are given, of the form $\widetilde{c}_{j_0,k} = c_{j_0,k} + e_k$, where the e_k are i.i.d. random variables. Although this might seem rather naive, it is close to the equidistant design case in a regression problem. Indeed, assuming for simplicity that f is defined on $[0, 1]$ and continuous at $x_k = 2^{-j_0}k$, $k = 1, \ldots, n = 2^{j_0}$, then

$$c_{j_0,k} = 2^{j_0/2} \int_{\mathbb{R}} \varphi(2^{j_0}x - k) f(x)\, dx \approx f(2^{-j_0}k) 2^{-j_0/2} = f(x_k) 2^{-j_0/2}, \tag{1.1}$$

since for the scaling function φ, $\int_{\mathbb{R}} \varphi(x)\, dx = 1$.

A word on notation is needed here: *Throughout the text*, $a \approx b$ *is used to mean* "*about the same*," *that is*, a/b *is approximately* 1, *and* $a_n \sim b_n$ *means that* a_n/b_n *tends to* 1 *as n tends to infinity.*

Assume now that we obtained (by some preprocessing) noisy observations of the $c_{j_0,k}$ $(:= f_k)$,

$$\widetilde{c}_{j_0,k} = f_k + e_k, \qquad k = 1, \ldots, 2^m = n, \tag{1.2}$$

where the e_k are i.i.d. random variables which represent the noise or the observation errors. Applying a fast wavelet transform W_n to the data $(\widetilde{c}_{j_0,k})$ gives

$$\widetilde{w}_{j,k} = (W_n(f))_{j,k} + (W_n(e))_{j,k} := w_{j,k} + z_{j,k}, \tag{1.3}$$

where for the transformed data, we write $w_{\cdot,\cdot}$ rather than $d_{\cdot,\cdot}$ since the coefficients do not exactly correspond to the wavelet coefficients, due to boundary effects and since we do not compute the whole "triangle" of coefficients but stop at some level j_{top} (and replace the top $d_{j,k}$, $j \geq j_{\text{top}}$, by the $c_{j_{\text{top}},k}$).

Again, since the wavelet transform of a "nice" function is sparse, it is expected that only a small fraction of the wavelet coefficients are big and that the rest are small and thus negligible. So if a $\widetilde{w}_{j,k}$ is small, it is reasonable to regard it as mostly noise and to set $w_{j,k}$ to zero; if it is big, it is reasonable to keep it. This is known as hard thresholding. Soft thresholding shrinks everything toward zero by a certain amount, thus reducing the variance of the estimation at the cost of a higher bias. Nonlinear shrinking policies were first applied to wavelet coefficients by Donoho and Johnstone [12–14, 16] and in a function space framework by DeVore and Lucier [11]. (This procedure has some origin in [17]; see also [5].) In both cases the threshold usually depends on the index (j, k). There is by now

a variety of papers on these rules, and other shrinking methods have also been proposed [6, 20] but, all in all, the asymptotic performances of the different shrinking estimators do not vary much. We refer the reader to the book of Vidakovic [38] for further information on wavelets in statistics.

Clearly, the thresholds should depend on the type of noise and on the variance of the noise in the initial data. For i.i.d. normal initial noise, the distribution of the noise in the wavelet coefficients is also i.i.d. normal. In practice, the normal noise assumption is not always realistic. For large datasets one can rely on the central limit theorem and get the same asymptotic results as in the normal case, but this will not do for small datasets (e.g., see [19]). Matters get worse if the requirement of independence in the initial noise is dropped; often the initial data to which the wavelet transform is applied is already the result of preprocessing (when dealing with irregularly spaced design, random design, density estimation), and then the noise is neither identically distributed nor independent. In particular, for noise with tails heavier than normal, the thresholds are sometimes too small. To date, only a few results directly deal with nonnormal noise ([19, 26, 27], etc.), and it is the purpose of the present work to help better understand such a case. Let us describe the content of the present paper. In the next section we recall the "ideal" denoising method and study it for certain classes of noise. Minimax type results are obtained in Section 3. In Section 4, compactly supported noises are tackled. Section 5 deals with compactly supported noise with a smooth density. In the last section, a different approach to choosing thresholds is introduced (the threshold of the kth coefficient is always the same, no matter the signal length). This is studied in a normal framework, the extension to nonnormal noise being briefly indicated. Various simulations and computations are also presented. At times, our approach also complements the Gaussian framework. Many of the results presented here have been announced in [2] and were presented at the 1996–1997 wavelet special year in Montréal. A companion paper [3] studies the function space approach to denoising in a not-necessarily normal framework.

2. The ideal method. Let us briefly recall Donoho and Johnstone's ideal denoising method [13]. Given noisy wavelet coefficients, that is, the true wavelet coefficient plus a random term which represents the noise and assuming that one has knowledge of the true wavelet coefficients, an ideal (oracular) estimator is to set a noisy coefficient to zero if the variance σ^2 of the noise is greater than the square of the true wavelet coefficient; otherwise the noisy coefficient is kept. The mean square error of this estimator is the minimum of σ^2 and of the square of the coefficient. Under the assumption of i.i.d. normal noise (see also [22] for normal correlated noise), these authors show that the soft thresholding estimator achieves a risk at most $O(\log n)$ times the risk of this ideal estimator. Moreover, no estimator is asymptotically better.

The "ideal method" does not require any a priori knowledge of the function to denoise, but might not be optimal when smoothness class information is available.

For many "smooth" functions "most of" the wavelet coefficients are rather small and only a small part of the wavelet coefficients are big. This means that the risk of the ideal estimator is small and this, in turn, implies that the risk of soft thresholding is small for these functions.

Let us now assume the regular design situation; that is, we have the observations $X_i = f_i + e_i$, $i = 1, \dots, n = 2^m$, where the e_i have finite variance. To this data we apply a discrete wavelet transform $W_n : \mathbb{R}^n \to \mathbb{R}^n$ which is adapted to an interval by boundary corrections or by periodization, in any case in such a way that W_n is an orthonormal transformation.

Let $Y = W_n(X)$ be the empirical wavelet coefficients, let $\theta = W_n(f)$ and let $z = W_n(e)$. Thus $Y_i = \theta_i + z_i$, $i = 1, \dots, n$, and the respective mean square errors in estimating θ and f are equal. Assuming some knowledge of the true wavelet coefficients θ, the oracular estimation for θ_i is then given by $\check{\theta}_i = Y_i$ if $\theta_i^2 > \sigma^2$ and $\check{\theta}_i = 0$ if $\theta_i^2 \le \sigma^2$. In plain words, an empirical wavelet coefficient is kept if its contribution to the energy of the function is greater than the variance of the noise; otherwise it is discarded. The performances of other estimators (in particular of the soft thresholding estimator $T_\lambda^S(x) = (|x| - \lambda)_+ \operatorname{sgn}(x)$ or of the hard thresholding estimator $T_\lambda^H(x) = x\mathbf{1}_{\{|x|>\lambda\}}$) when applied to Y are compared to the benchmark

$$B_n(\theta, \sigma^2) := \sigma^2 + \sum_{i=1}^{n} \min(\theta_i^2, \sigma^2), \tag{2.1}$$

which is the mean square error of $\check{\theta}$ plus σ^2 which itself represents the penalty corresponding to the error of the oracular estimator $\check{\theta}$ if $\theta = 0$ (see [13]). This is close to assuming that at least one θ_i^2 is greater than σ^2; note also that $B_n(\theta, \sigma^2)$ is small in comparison to $n\sigma^2$ (the sum of the variances of the components of X) if θ (the wavelet transform) is sparse.

A further note is necessary here: let f be a function defined on $[0, 1]$ and let W_n be a wavelet transform based on the wavelet ψ; then it follows from (1.1) that

$$W_n\big(f(i/n)_{i=1,\dots,n}\big)_{j,k} \sim \sqrt{n}\langle f, \psi_{j,k}\rangle.$$

So, in the sequel, when the thresholds are increasing with n, it is important to remember that the true wavelet coefficients are also increasing with n, whereas the variance of the noise in the coefficients remains constant.

If the e_i are i.i.d. normal random variables, then so are the z_i. Now, for $Y_i = \theta_i + z_i$, $i = 1, \dots, n$, $n \ge 3$, where the θ_i are parameters of interest and where the z_i are centered i.i.d. normal random variables with variance σ^2, Donoho and Johnstone [13] proved that

$$\sup_{\theta\in\mathbb{R}^n} \frac{E\|T_{\lambda_n^*}^S(Y) - \theta\|^2}{B_n(\theta, \sigma^2)} \le \Lambda_n^* := \inf_\lambda \sup_{t\in\mathbb{R}} \frac{E(T_\lambda^S(X) - t)^2}{n^{-1} + \min(t^2, 1)} \le (1 + 2\log n),$$

where $\lambda_n^* \le \sigma\sqrt{2\log n}$ is the largest λ attaining Λ_n^* and where $X \sim N(t, 1)$. (Here and throughout, $\|\cdot\|$ denotes the Euclidean norm and θ_i are real parameters of

interest.) A result of this type is also proved for i.i.d. random variables with exponential tails in [19].

Inspired by the above results and their proofs, we now provide for i.i.d. random variables z_i with (known) distribution, a general (nonasymptotic) estimate on the ratio of thresholding risk and benchmark. It should also be emphasized here that the result below (as well as Proposition 2.5) does not depend at all on the correlation among the coefficients (except for their variances). However, for highly correlated coefficients, the benchmark as defined by (2.1) might not be such a good measure of estimation.

THEOREM 2.1. *Let $Y_i = \theta_i + z_i$, $i = 1, \dots, n$, $n \geq 4$, where the z_i are identically distributed symmetric (about zero) random variables with law μ and such that $Ez_1^2 = \sigma^2$. Then, the equation*

$$(n+1)p(\lambda, 0) := 2(n+1)\int_\lambda^\infty (x-\lambda)^2 \mu(dx) = \lambda^2 + \sigma^2, \tag{2.2}$$

has a unique positive solution λ_n. Moreover,

$$\Lambda_n := \inf_\lambda \sup_{\theta \in \mathbb{R}^n} \frac{E\|T_\lambda^S(Y) - \theta\|^2}{B_n(\theta, \sigma^2)} = \sup_{\theta \in \mathbb{R}^n} \frac{E\|T_{\lambda_n}^S(Y) - \theta\|^2}{B_n(\theta, \sigma^2)} = \frac{n(\lambda_n^2 + \sigma^2)}{(n+1)\sigma^2}.$$

PROOF. For $\lambda \geq 0, \theta \in \mathbb{R}$, let

$$\begin{aligned} p(\lambda, \theta) &:= E|T_\lambda^S(z_1 + \theta) - \theta|^2 \\ &= \int_{\mathbb{R}} \big(\operatorname{sgn}(x+\theta)(|x+\theta| - \lambda)_+ - \theta\big)^2 \mu(dx), \end{aligned} \tag{2.3}$$

and let also

$$L_n(\lambda) := \sup_{\theta \in \mathbb{R}} \frac{p(\lambda, \theta)}{\sigma^2/n + \min(\theta^2, \sigma^2)}.$$

Then

$$\sum_{i=1}^n E|T_\lambda^S(Y_i) - \theta_i|^2 \leq L_n(\lambda) \sum_{i=1}^n \left(\frac{\sigma^2}{n} + \min(\theta_i^2, \sigma^2) \right) = L_n(\lambda) B_n(\theta, \sigma^2). \tag{2.4}$$

The function $p(\lambda, \infty) := \lim_{\theta \to \infty} p(\lambda, \theta) = \sigma^2 + \lambda^2$ is continuous and increasing on $[0, \infty)$, whereas $p(\lambda, 0) = 2\int_\lambda^\infty (x-\lambda)^2 \mu(dx)$ is continuous and nonincreasing on $[0, \infty)$ (decreasing on the positive part of the support of μ, and zero outside the support). Moreover, $p(0, 0) = p(0, \infty) = \sigma^2$. Hence λ_n, which is the unique solution of

$$\frac{p(\lambda, 0)}{\sigma^2/n} = \frac{p(\lambda, \infty)}{\sigma^2 + \sigma^2/n},$$

minimizes

$$\sup_{\theta\in\{0,\infty\}} \frac{p(\lambda,\theta)}{\sigma^2/n+\min(\theta^2,\sigma^2)},$$

that is,

$$\begin{aligned}\frac{n(\lambda_n^2+\sigma^2)}{(n+1)\sigma^2} &= \inf_{\lambda}\sup_{\theta\in\{0,+\infty\}} \frac{p(\lambda,\theta)}{\sigma^2/n+\min(\theta^2,\sigma^2)} \\ &= \sup_{\theta\in\{0,+\infty\}} \frac{p(\lambda_n,\theta)}{\sigma^2/n+\min(\theta^2,\sigma^2)}.\end{aligned} \tag{2.5}$$

We now claim that

$$L_n(\lambda_n) = \sup_{\theta\in\mathbb{R}} \frac{p(\lambda_n,\theta)}{\sigma^2/n+\min(\theta^2,\sigma^2)} = \sup_{\theta\in\{0,\infty\}} \frac{p(\lambda_n,\theta)}{\sigma^2/n+\min(\theta^2,\sigma^2)}. \tag{2.6}$$

Let λ be fixed, let δ_θ be the Dirac measure with unit mass at θ and let

$$f_{\lambda,\theta}(x) := (T_\lambda^S(x)-\theta)^2 = \begin{cases}(x-\theta+\lambda)^2, & x\in(-\infty,-\lambda],\\ \theta^2, & x\in(-\lambda,\lambda),\\ (x-\theta-\lambda)^2, & x\in[\lambda,\infty).\end{cases}$$

If $\lambda > h > 0$, $\theta > 0$, then

$$\begin{aligned}p(\lambda,\theta+h) &= \int_{\mathbb{R}} f_{\lambda,\theta+h}(x)(\mu*\delta_{\theta+h})(dx)\\ &= \int_{\mathbb{R}} f_{\lambda,\theta+h}(x+h)(\mu*\delta_{\theta})(dx)\\ &\geq \int_{\mathbb{R}} f_{\lambda,\theta}(x)(\mu*\delta_{\theta})(dx) = p(\lambda,\theta),\end{aligned}$$

where the inequality holds since

$$\begin{aligned}f_{\lambda,\theta+h}(x+h) &= \begin{cases}(x-\theta+\lambda)^2, & x\in(-\infty,-\lambda-h],\\ (\theta+h)^2, & x\in(-\lambda-h,\lambda-h),\\ (x-\theta-\lambda)^2, & x\in[\lambda-h,\infty),\end{cases}\\ &\geq f_{\lambda,\theta}(x).\end{aligned}$$

Thus $p(\lambda,\theta)$ is nondecreasing in θ on $(0,\infty)$ and nonincreasing on $(-\infty,0)$ since μ is symmetric. Therefore

$$\sup_{|\theta|\geq\sigma} \frac{p(\lambda,\theta)}{\sigma^2/n+\min(\theta^2,\sigma^2)} = \frac{p(\lambda,\infty)}{\sigma^2/n+\sigma^2}\left(=\frac{p(\lambda,-\infty)}{\sigma^2/n+\sigma^2}\right). \tag{2.7}$$

We now claim that $\theta^2 + p(\lambda, 0) \geq p(\lambda, \theta)$. Indeed,

$$\theta^2 + f_{\lambda,0}(x) = \begin{cases} \theta^2 + (x+\lambda)^2, & x \in (-\infty, -\lambda], \\ \theta^2, & x \in (-\lambda, \lambda), \\ \theta^2 + (x-\lambda)^2, & x \in [\lambda, \infty), \end{cases}$$

$$\geq \begin{cases} (x+\lambda)^2, & x \in (-\infty, -\lambda - \theta], \\ \theta^2, & x \in (-\lambda - \theta, \lambda - \theta), \\ (x-\lambda)^2, & x \in [\lambda - \theta, \infty), \end{cases}$$

$$= f_{\lambda,\theta}(x + \theta).$$

Hence

$$\begin{aligned} \theta^2 + p(\lambda, 0) &= \int_{\mathbb{R}} \big(\theta^2 + f_{\lambda,0}(x)\big)\mu(dx) \geq \int_{\mathbb{R}} f_{\lambda,\theta}(x+\theta)\mu(dx) \\ &= \int_{\mathbb{R}} f_{\lambda,\theta}(x)(\mu * \delta_\theta)(dx) = p(\lambda, \theta), \end{aligned} \tag{2.8}$$

which proves our claim. Assume for a moment that $p(\lambda_n, 0) \geq \sigma^2/n$; it then follows from (2.8) that for $\theta \in [-\sigma, \sigma]$,

$$\frac{p(\lambda_n, \theta)}{\sigma^2/n + \min(\theta^2, \sigma^2)} = \frac{p(\lambda_n, \theta)}{\sigma^2/n + \theta^2} \leq \frac{\theta^2 + p(\lambda_n, 0)}{\sigma^2/n + \theta^2} \leq \frac{p(\lambda_n, 0)}{\sigma^2/n}. \tag{2.9}$$

We therefore conclude from (2.7) and (2.9) that if $p(\lambda_n, 0) \geq \sigma^2/n$, then

$$\sup_{\theta \in \mathbb{R}} \frac{p(\lambda_n, \theta)}{\sigma^2/n + \min(\theta^2, \sigma^2)} = \sup_{\theta \in \{0, \infty\}} \frac{p(\lambda_n, \theta)}{\sigma^2/n + \min(\theta^2, \sigma^2)}.$$

Let us now show for $n \geq 4$, $p(\lambda_n, 0) \geq \sigma^2/n$, or equivalently, since $p(\lambda_n, 0)/(\sigma^2/n) = (\sigma^2 + \lambda_n^2)/(1 + 1/n)\sigma^2$, that $\lambda_n^2 \geq \sigma^2/n$. Indeed,

$$p(\lambda, 0) = Eg(z_1^2) \geq g(Ez_1^2) = g(\sigma^2) = Eg(Y^2) = p_Y(\lambda, 0),$$

where $g(x) = (\sqrt{|x|} - \lambda)_+^2$ is convex, where Y is a random variable with law $(\delta_{-\sigma} + \delta_{+\sigma})/2$ and where $p_Y(\lambda, 0)$ is as in (2.2) with μ replaced by $(\delta_{-\sigma} + \delta_{+\sigma})/2$. Hence $p(\lambda, 0) \geq p_Y(\lambda, 0)$ which implies that $\lambda_n \geq \xi_n$, where ξ_n is the solution of $(n+1)p_Y(\xi, 0) = \xi^2 + \sigma^2$. Thus it is enough to prove that for $n \geq 4$, $\xi_n^2 \geq \sigma^2/n$, but this is easily verified since ξ_n is the solution (smaller or equal to σ) of $(n+1)(\sigma - \xi)^2 = \xi^2 + \sigma^2$. Hence (2.4)–(2.6) show that

$$\Lambda_n \leq \sup_{\theta \in \mathbb{R}^n} \frac{E\|T_{\lambda_n}^S(Y) - \theta\|^2}{B_n(\theta, \sigma^2)} \leq L_n(\lambda_n) = \frac{n(\lambda_n^2 + \sigma^2)}{(n+1)\sigma^2}.$$

To finish the proof, we now show that λ_n is the optimal threshold. For $\lambda > 0$ we have (choosing $\theta = 0$ or $\theta = \infty$)

$$\sup_{\theta \in \mathbb{R}^n} \frac{E\|T_\lambda^S(Y) - \theta\|^2}{B_n(\theta, \sigma^2)} \geq \max\left(\frac{np(\lambda, 0)}{\sigma^2}, \frac{np(\lambda, \infty)}{(n+1)\sigma^2}\right) \geq L_n(\lambda_n),$$

where the second inequality holds since λ_n minimizes the second term and $L_n(\lambda_n)$ is the minimum of this term by (2.5) and (2.6). This finishes the proof. $\square$

REMARK 2.2. If the z_i are compactly supported on $[-M, M]$, $\lim_{\lambda\to M} p(\lambda, 0) = 0$, thus $\lim_{n\to+\infty} \lambda_n = M$ and $\lim_{n\to+\infty} \Lambda_n = (\sigma^2 + M^2)/\sigma^2$. In this case the risk associated with soft thresholding is comparable to the benchmark. For example, if the law of z_1 is $(\delta_{-\sigma} + \delta_{+\sigma})/2$, then as $n \to +\infty$, $\lambda_n \to \sigma$ and $\Lambda_n \to 2$. Thus the soft thresholding risk is only twice as bad as the benchmark. For the uniform distribution on $[-M, M]$, the ratio is 4. The ratio becomes worse if the size of the support is large in comparison to the variance.

REMARK 2.3. Theorem 2.1 is still true if μ is not symmetric. It was notationally convenient not to include this case above. Replacing in (2.2),

$$2\int_\lambda^\infty (x-\lambda)^2\mu(dx) \qquad \text{by } \int_{|x|>\lambda}^\infty (|x|-\lambda)^2\mu(dx),$$

gives a result for arbitrary μ. However, it might be better to use different thresholds for each side, but this leads to another class of estimators. Nevertheless, then the optimal pair of thresholds $(\lambda_{\text{left}}, \lambda_{\text{right}})$ is the solution of

$$\begin{aligned}(n+1)&\left(\int_{-\infty}^{\lambda_{\text{left}}} (|x|-\lambda_{\text{left}})^2\mu(dx) + \int_{\lambda_{\text{right}}}^\infty (|x|-\lambda_{\text{right}})^2\mu(dx)\right)\\ &= \max(\lambda_{\text{left}}^2, \lambda_{\text{right}}^2) + \sigma^2,\end{aligned}$$

as easily seen by the methods presented above.

REMARK 2.4. The previous method of proof can also be applied to other loss functions, thus removing the second moment assumption on the z_i in Theorem 2.1. This is briefly explained now. Let ℓ be an even convex function with $\ell(0) = 0$, and let $m_\ell := E\ell(z_1) < +\infty$. Let $B_n(\theta, m_\ell) := m_\ell + \sum_{i=1}^n \min(\ell(\theta_i), m_\ell)$ and finally let $p_\ell(\lambda, \theta) := E\ell(T_\lambda^S(z_1 + \theta) - \theta)$. Then, for all $n \ge n_0$, the equation

$$(n+1)p_\ell(\lambda, 0) = 2(n+1)\int_\lambda^\infty \ell(x-\lambda)\mu(dx) = E\ell(z_1 - \lambda) \tag{2.10}$$

has a unique positive solution λ_n with, moreover,

$$\sup_{\theta\in\mathbb{R}^n} \frac{E\sum_{i=1}^n \ell(T_{\lambda_n}^S(Y_i) - \theta_i)}{B_n(\theta, m_\ell)} = \inf_\lambda \sup_{\theta\in\mathbb{R}^n} \frac{E\sum_{i=1}^n \ell(T_\lambda^S(Y_i) - \theta_i)}{B_n(\theta, m_\ell)} = \frac{nE\ell(z_1 - \lambda_n)}{(n+1)m_\ell}.$$

As in the previous proof, we see that for $\lambda \ge 0$, $p_\ell(\lambda, 0)$ is continuous and nonincreasing (decreasing on the positive part of the support of μ and zero outside) while (by the convexity of ℓ and since μ is symmetric) $p_\ell(\lambda, \infty) = E\ell(z_1 - \lambda)$ is continuous and nondecreasing and moreover $p_\ell(0, \infty) = m_\ell$. These ensure that

$$\frac{p_\ell(\lambda, 0)}{m_\ell/n} = \frac{p_\ell(\lambda, \infty)}{m_\ell + m_\ell/n},$$

has a unique positive solution λ_n which minimizes

$$\sup_{\theta\in\{0,\infty\}} \frac{p_\ell(\lambda,\theta)}{m_\ell/n+\min(\ell(\theta),m_\ell)},$$

and $np_\ell(\lambda_n,\infty)/(n+1)m_\ell$ is equal to this minimum. This leads us to a claim similar to (2.6) but with p replaced by p_ℓ and σ^2 replaced by m_ℓ. To prove it, proceeding as above, we arrive at

$$\sup_{|\theta|\geq\ell^{-1}(m_\ell)} \frac{p_\ell(\lambda,\theta)}{m_\ell/n+\min(\ell(\theta),m_\ell)} = \frac{p_\ell(\lambda,\infty)}{m_\ell/n+m_\ell}\left(=\frac{p_\ell(\lambda,-\infty)}{m_\ell/n+m_\ell}\right), \tag{2.11}$$

which is the ℓ-version of (2.7). Then for all θ, $m_\ell + p_\ell(\lambda,0) \geq p_\ell(\lambda,\theta)$ and assuming for a moment that $np_\ell(\lambda_n,0)\geq m_\ell$, it then follows that for $\theta\in[-\ell^{-1}(m_\ell),\ell^{-1}(m_\ell)]$,

$$\frac{p_\ell(\lambda_n,\theta)}{m_\ell/n+\min(\ell(\theta),m_\ell)} = \frac{p_\ell(\lambda_n,0)}{m_\ell/n}.$$

This proves that for all n such that $np_\ell(\lambda_n,0)\geq m_\ell$,

$$\sup_{\theta\in\mathbb{R}} \frac{p_\ell(\lambda_n,\theta)}{m_\ell/n+\min(\ell(\theta),m_\ell)} = \sup_{\theta\in\{0,\infty\}} \frac{p_\ell(\lambda_n,\theta)}{m_\ell/n+\min(\ell(\theta),m_\ell)},$$

and the result is proved.

Unfortunately the previous theorem does not, in general, directly apply to the noisy wavelet coefficients; the requirement of identically distributed random variables is too strong. In the case of nonidentically distributed random variables, the next result gives a good suggestion for a threshold and an upper bound for the ratio of risks, namely compute $\lambda_{n,i}$ for each z_i separately, and choose for global threshold the largest of those $\lambda_n = \sup_i \lambda_{n,i}$. This will give a bound on the ratio of risks which is more handy (although bigger) than $\sup_i \Lambda_{n,i}$ (note that $\inf_i \Lambda_{n,i}$ provides a lower bound). Note also that below the benchmark $B_n(\theta,\sigma)$ has been replaced by $\sum_{i=1}^n(\min(\theta_i^2,\sigma_i^2)+\sigma_i^2/n)$ since the variances σ_i^2 are no longer the same and the oracular estimation is done coefficient by coefficient. In light of the remarks above, it is also clear that below the symmetry (or quadratic loss) assumption can be removed.

PROPOSITION 2.5. *Let $Y_i=\theta_i+z_i$, $i=1,\dots,n$ with $n\geq 4$, where the z_i are symmetric (about zero) random variables with law μ_i such that $Ez_i^2=\sigma_i^2$. Let $\bar\sigma^2=(\sum_{i=1}^n\sigma_i^2)/n$. For each i, let $\lambda_{n,i}$ be the unique solution of the equation*

$$2(n+1)\int_\lambda^\infty (x-\lambda)^2\mu_i(dx) = \lambda^2+\sigma_i^2, \qquad \lambda>0, \tag{2.12}$$

let $\lambda_n \geq \sup_i \lambda_{n,i}$ *and let* $\Lambda_n := \sup_i (\lambda_n^2 + \sigma_i^2)/(\sigma_i^2(1+1/n))$. *Then*

$$\sup_{\theta \in \mathbb{R}^n} \frac{E\|T_{\lambda_n}^S(Y) - \theta\|^2}{\bar{\sigma}^2 + \sum_{i=1}^n \min(\theta_i^2, \sigma_i^2)} \leq \Lambda_n. \tag{2.13}$$

PROOF. The proof is similar to the proof of the previous theorem. All we need to show is that

$$\sup_{\theta \in \mathbb{R}} \frac{p_i(\lambda_n, \theta)}{\sigma_i^2/n + \min(\theta^2, \sigma_i^2)} \leq \Lambda_n, \qquad i = 1, \ldots, n, \tag{2.14}$$

where $p_i(\lambda, \theta) := \int_{\mathbb{R}} |T_\lambda^S(x+\theta) - \theta|^2 \mu_i(dx)$. As shown in the previous proof, the left-hand side in (2.14) is smaller than the maximum of

$$\frac{\lambda_n^2 + \sigma_i^2}{(1+1/n)\sigma_i^2} \quad \text{and} \quad \sup_{\theta \in [-\sigma_i, \sigma_i]} \frac{p_i(\lambda_n, \theta)}{\sigma_i^2/n + \theta^2}.$$

Now, the first term above is dominated by Λ_n while for the second term,

$$\begin{aligned} \sup_{\theta \in [-\sigma_i, \sigma_i]} \frac{p_i(\lambda_n, \theta)}{\sigma_i^2/n + \theta^2} &\leq \sup_{\theta \in [-\sigma_i, \sigma_i]} \frac{p_i(\lambda_n, 0) + \theta^2}{\sigma_i^2/n + \theta^2} \leq \sup_{\theta \in [-\sigma_i, \sigma_i]} \frac{p_i(\lambda_{n,i}, 0) + \theta^2}{\sigma_i^2/n + \theta^2} \\ &= \frac{\lambda_{n,i}^2 + \sigma_i^2}{\sigma_i^2(1+1/n)} \leq \Lambda_n, \end{aligned}$$

where the second and third inequalities hold since $\lambda_{n,i} \leq \lambda_n$, and the equality holds because of the properties of $\lambda_{n,i}$ obtained in the proof of Theorem 2.1. □

We want to apply the above theorem to empirical wavelet coefficients $Y_k = \theta_k + z_k$. Since W_n is orthonormal, the z_k are linear combinations of the initial noise, that is, $z_k = \sum_{i=1}^n w_{k,i}^{(n)} e_i$ with $\sum_{i=1}^n (w_{k,i}^{(n)})^2 = 1$ and $\lambda_{n,k}$ is thus quite complicated to compute. An alternative is to find an upper bound for the $\lambda_{n,k}$ which only depends on the initial noise (e_i). By Proposition 2.5, to compute an upper bound for the performance of soft thresholding, we just need to find an upper bound for all the optimal thresholds. Now, for given symmetric distributions μ_1 and μ_2 with $\mu_1([t, \infty)) \leq \mu_2([t, \infty))$, $t \geq 0$, and $\sigma_1^2 = \sigma_2^2 = \sigma^2$, let λ_i, $i = 1, 2$, be the solutions of

$$2(n+1) \int_\lambda^\infty (x - \lambda)^2 \mu_i(dx) = \lambda^2 + \sigma^2, \qquad \lambda > 0,\ n \in \mathbb{N},\ i = 1, 2;$$

where clearly $\lambda_2 \geq \lambda_1$. By a classical result of Hoeffding (see [35], page 855), if the e_i are zero mean, bounded i.i.d. random variables (or just bounded independent or even bounded martingale differences) and, say, with support $[-M, M]$,

$$P(z_k \geq t) \leq \exp\left(\frac{-2t^2}{\sum_i w_{k,i}^2 (M - (-M))^2}\right) = \exp\left(\frac{-t^2}{2M^2}\right)$$

for $t \geq 0$. Thus an upper bound for the optimal thresholds of the empirical wavelet coefficients is the solution of (2.2) for the symmetric distribution μ_2 defined by $\mu_2([t,\infty)) := \exp(-t^2/(2M^2))$. In Theorem 2.7, we show that for this μ_2, the solution is asymptotically like $\sqrt{2M^2 \log n}$. Estimates more precise than $\exp(-t^2/(2M^2))$ are available but the asymptotics of the thresholds do not change much with these. For centered (e_i) with exponential tails, estimates on $P(z_k \geq t)$ are known ([34], Section 2.2) and these also provide upper bounds for the optimal thresholds for this class of noise.

If (e_i) has tails heavier than normal, then the central limit theorem applies and the distribution of $\sum_i w_{k,i} e_i$ "tends" to the normal distribution; that is, its tail is becoming smaller. This tail compression property and the fact that the family of variance mixtures of normal random variables is closed under mixtures and convolutions leads us to the following which provides another approach to finding threshold upper bounds.

PROPOSITION 2.6. *Let μ be a variance mixture of normal distributions, that is, the measure μ defined on $\mathbb{R}$ is absolutely continuous with density $\int_0^\infty \phi_s(x)\nu(ds)$, where ν is a probability measure on $(0,\infty)$ and where ϕ_s is the centered normal density of variance s. Let X_i, $i = 1,\dots,n$, be i.i.d. random variables with law μ, and let g be a convex function on $\mathbb{R}^+$. Finally, let $a_1,\dots,a_n \in \mathbb{R}$, $b_1,\dots,b_n \in \mathbb{R}$, with respective squares written in nonincreasing order $(\cdots a_i^2 \geq a_{i+1}^2 \geq \cdots)$, $(\cdots b_i^2 \geq b_{i+1}^2 \geq \cdots)$ be such that $\sum_{i=1}^k a_i^2 \leq \sum_{i=1}^k b_i^2$, $k = 1,\dots,n-1$, and $\sum_{i=1}^n a_i^2 = \sum_{i=1}^n b_i^2$. Then,*

$$Eg\left(\left(\sum_{i=1}^n a_i X_i\right)^2\right) \leq Eg\left(\left(\sum_{i=1}^n b_i X_i\right)^2\right).$$

PROOF. For $y \geq 0$, let $R^N(y) := Eg(yN^2)$ where N is a standard normal random variable. Clearly, R^N is a convex function since $y \to g(yx^2)$ is convex for all $x \in \mathbb{R}$. Next, it is easy to see that, for any $a_1,\dots,a_n \in \mathbb{R}$ and X_i i.i.d. with law a variance mixture of normals, $\sum_{i=1}^n a_i X_i$ is again a variance mixture of normal random variables. The mixing measure of this sum is given by $\nu_{a_1^2} * \cdots * \nu_{a_n^2}$, where for any $c > 0$, ν_c is the measure $\nu(\cdot/c)/c$ and ν_0 is the Dirac measure with unit mass at 0, and where ν is the mixing measure of X_i. Now,

$$\begin{aligned} R(a_1^2,\dots,a_n^2) &:= Eg\left(\left(\sum_{i=1}^n a_i X_i\right)^2\right) \\ &= \int_{-\infty}^{+\infty} g(x^2)\left(\int_0^\infty \cdots \int_0^\infty \phi_{\sum_{i=1}^n a_i^2 u_i}(x)\nu(du_1)\cdots\nu(du_n)\right)dx \end{aligned}$$

$$= \int_0^\infty \cdots \int_0^\infty \left(\int_{-\infty}^{+\infty} g(x^2) \phi_{\sum a_i^2 u_i}(x)\, dx \right) \nu(du_1) \cdots \nu(du_n)$$

$$= E R^N \left(\sum_{i=1}^n a_i^2 Y_i \right),$$

where the Y_i are i.i.d. random variables with law ν. Since R^N is convex, it follows from [30], Chapter 12, beginning of Section G that R is Schur-convex, which gives the result. □

Let us now briefly explain how the previous proposition can be used when the initial noise $e = (e_i)$ is i.i.d. with law μ a variance mixture of normals (see also Proposition 1.5 in [2]). For $x \geq 0$, let $g(x) = (\sqrt{x} - \lambda)_+^2$. Clearly, g is convex and $g(x^2) = (|x| - \lambda)_+^2$. Now let $b_1 = 1,\ b_2 = \cdots = b_n = 0$ and let $z = (z_i)_i$ be the wavelet transform of e, that is, $z_i = (W_n e)_i = \sum_{j=1}^n w_{i,j}^{(n)} e_j$, where without loss of generality we assume that the $(w_{i,j}^{(n)})^2$ are in nonincreasing order (for the index j). Since $\sum_{j=1}^n (w_{i,j}^{(n)})^2 = 1$, then

$$E\left(\left| \sum_j w_{i,j}^{(n)} e_j \right| - \lambda \right)_+^2 \leq E\left(\left| \sum_j b_j e_j \right| - \lambda \right)_+^2,$$

for all $\lambda > 0$. But the terms above are, respectively, equal to $2\int_\lambda^\infty (x - \lambda)^2 \mu_i(dx)$ and $2\int_\lambda^\infty (x - \lambda)^2 \mu(dx)$, where μ_i is the law of z_i. This implies that the solution of (2.2) is larger than the solution $\lambda_{n,i}$ of (2.12), providing via (2.13) an upper bound on the ratio of risks.

The class of variance mixtures of the normal distribution contains many important distributions, for example, densities of the form $h(x) = c_1 \exp(-c_2 x^d)$, $0 < d < 2$, or $c_3/(1 + x^2)^n,\ n \geq 1$, where c_1, c_2, c_3 are appropriate constants (see [18], Chapter XIII and [2]). However, for a specific application, this class might not be adequate. Nevertheless, it is sometimes possible to carry over their properties to other "nearby" densities, since we only need an upper bound on $p(\lambda, 0)$. Indeed, if e_i is symmetric, and if there exists a random variable v whose distribution is a normal mixture such that $P(|e_1| > x) \leq K P(|v| > x)$, for all $x \geq 0$ and some $K \geq 1$, then for any $a_1, a_2, \ldots, \in \mathbb{R}$,

$$E\left(T_\lambda^S \left(\left| \sum_i a_i e_i \right| \right) \right)^2 \leq E\left(T_\lambda^S \left(K \left| \sum_i a_i v_i \right| \right) \right)^2,$$

where the v_i are i.i.d. copies of v (this last inequality is a consequence of the contraction principle (e.g., see [29], Lemma 4.6). Next, if λ_n is the positive solution of $(n + 1)E(|Kv| - \lambda)_+^2 = \lambda^2 + \sigma^2$, then

$$\sup_{\theta \in \mathbb{R}} \frac{E(T_{\lambda_n}^S(\sum_i a_i e_i + \theta) - \theta)^2}{\sigma^2/n + \min(\theta, \sigma^2)} \leq \Lambda_n, \tag{2.15}$$

where now the coefficients $a_\cdot$ are given by the entries of W_n and where $\Lambda_n = (\lambda_n^2 + \sigma^2)/((1 + 1/n)\sigma^2)$. The inequality (2.15) then implies a version like (2.12), for $\theta \in \mathbb{R}^n$. If $P(|e_1| > x) \le K P(|v| > x)$ holds only for $x \ge x_0$, then (again by [29], Section 4.2),

$$
\begin{aligned}
&E\left(T_\lambda^S\left(\left|\sum_i a_i e_i\right|\right)\right)^2 \\
&\quad \le \tfrac{1}{2} E\left(T_\lambda^S\left(2K x_0 \left|\sum_i a_i v_i\right|\right)\right)^2 + \tfrac{1}{2} E\left(T_\lambda^S\left(2K\left|\sum_i a_i u_i\right|\right)\right)^2 \\
&\quad := k(\lambda)
\end{aligned} \tag{2.16}
$$

where the u_i are i.i.d. random variables with distribution $(\delta_1 + \delta_{-1})/2$ and again the $a_\cdot$ are given by the entries of W_n. So if λ_n is the solution of $(n+1)k(\lambda) = \lambda^2 + \sigma^2$, then again (2.15) holds with $\Lambda_n = (\lambda_n^2 + \sigma^2)/((1 + 1/n)\sigma^2)$.

It would be desirable to generalize Proposition 2.6 to a wider class of distributions, or to obtain a version which directly describes the behavior of the tails of the sums. Results of this type are available for special classes of distributions, where the tail of the sums is bounded by the tail of the initial random variable [4].

In general, the distribution of the noise in the coarser levels of the transformed signal is difficult to compute, and as a consequence of Propositions 2.5 and 2.6, we used thresholds based on the initial noise. If this noise has tails heavier than normal, then the thresholds for the coarser levels are higher than actually necessary. In a minimax approach this does not really matter. As shown in [3], and depending on the noise, from some level on upward the thresholds can be chosen as in the normal case. Also for the finer level, the thresholds based on the initial noise are also often too high. For example, let X be a random variable with a Laplace distribution with $EX^2 = 1$ and λ_n such that $(1+n)p_X(\lambda, 0) = \lambda^2 + 1$. Let Y be an independent copy of X and let ξ_n be the solution of $(1+n)p_{(X+Y)/\sqrt{2}}(\xi, 0) = \xi^2 + 1$ [again, p_X and $p_{(X+Y)/\sqrt{2}}$ are defined as in (2.3) with μ, respectively, replaced by the law of X and the law of $(X+Y)/\sqrt{2}$]. Then (see Figure 1 and thanks to G. Nason) for $n = 512$, $\lambda_n = 2.85$, for $n = 2^{16}$, $\lambda_n = 5.43$, while the respective values of ξ_n are $\xi_n = 1.81$ and $\xi_n = 3.55$. So, for the Laplace distribution, it is reasonable to expect that the optimal thresholds for the finer level wavelet coefficients are much smaller than the optimal ones.

To finish this section, asymptotic rates for λ_n and Λ_n are obtained for exponentially decreasing densities. This result applies when h, below, grows asymptotically at least as fast as a fractional polynomial and, in particular, taking $h(x) = x^2/2$ recovers a normal result given in [13]. Some numerical comparisons between optimal and asymptotic thresholds for some classes of distributions are also presented.

THEOREM 2.7. *Let μ be absolutely continuous with variance σ^2 and density $f(x) = C\exp(-h(x))$ (C normalizing constant) where h is even, continuous and increasing on $[0,\infty)$, and such that $\liminf_{x\to\infty} \frac{h(cx)}{h(x)} > 1$, for all $c > 1$. Let λ_n be the solution of (2.2) and let $\Lambda_n = n(\lambda_n^2 + \sigma^2)/(\sigma^2(n+1))$. Then*

$$\lim_{n\to\infty} \frac{h^{-1}(\log n)}{\lambda_n} = 1 \quad \textit{and} \quad \lim_{n\to\infty} \frac{(h^{-1}(\log n))^2 + \sigma^2}{\sigma^2 \Lambda_n} = 1.$$

PROOF. First let us note a consequence of the conditions imposed on h [showing, e.g., that in the proof we only need $f(x) = C\exp(-h(x))$, for x large enough]. If λ tends to infinity, then for all $x \in \mathbb{R}$, $h(c\lambda) - h(x+\lambda) - \log(\lambda^2+\sigma^2)$ tends to plus infinity if $c > 1$ and to minus infinity if $0 < c < 1$. Set $q(\lambda) := (\lambda^2+\sigma^2)/2\int_\lambda^\infty (x-\lambda)^2 f(x)\,dx - 1, \lambda > 0$, which clearly defines an increasing function. For $0 < \delta < 1$, we now claim that $q(\lambda) f((1+\delta)\lambda) \to 0$, and that $q(\lambda) f((1-\delta)\lambda) \to +\infty$, as $\lambda \to +\infty$. For the first limit, let $\lambda \geq 1$. Then

$$\begin{aligned}
&\left(\frac{\lambda^2+\sigma^2}{2\int_\lambda^\infty (x-\lambda)^2 f(x)\,dx} - 1\right) f\big((1+\delta)\lambda\big) \\
&\quad \leq \frac{\lambda^2+\sigma^2}{\int_\lambda^\infty (x-\lambda)^2 f(x)\,dx} f\big((1+\delta)\lambda\big) \\
&\quad = \left(\int_0^\infty x^2 \frac{f(x+\lambda)}{f((1+\delta)\lambda)(\lambda^2+\sigma^2)}\,dx\right)^{-1} \to 0,
\end{aligned}$$

since

$$\frac{f(x+\lambda)}{f((1+\delta)\lambda)(\lambda^2+\sigma^2)} = \exp\big(h\big((1+\delta)\lambda\big) - \log(\lambda^2+\sigma^2) - h(x+\lambda)\big)$$

and, by assumption, this last term tends to infinity, for all x, as $\lambda \to +\infty$. Let us turn to the second limit,

$$\begin{aligned}
q(\lambda) f\big((1-\delta)\lambda\big) \;&\geq\; \left(2\int_0^\infty x^2 \frac{f(x+\lambda)}{f((1-\delta)\lambda)(\lambda^2+\sigma^2)}\,dx\right)^{-1} - f\big((1-\delta)\lambda\big) \\
&\to +\infty,
\end{aligned}$$

as $\lambda \to +\infty$. The second summand on the right-hand side above converges to 0. The integrand in the first summand is equal to $x^2\exp(h((1-\delta)\lambda) - h(x+\lambda) - \log(\lambda^2+\sigma^2))$, which, by assumption, converges to zero, for all $x \in \mathbb{R}$. Therefore the integral tends to 0 and the whole term to plus infinity. Thus for x sufficiently large (depending on δ),

$$\frac{1}{f((1-\delta)x)} \leq q(x) \leq \frac{1}{f((1+\delta)x)},$$

hence

$$\frac{1}{1-\delta} f^{-1}(1/y) \geq q^{-1}(y) \geq \frac{1}{1+\delta} f^{-1}(1/y),$$

for y large enough. Since λ_n is the solution of $q(\lambda_n) = n$, it follows that

$$\lim_{n\to\infty} \frac{\lambda_n}{f^{-1}(1/n)} = \lim_{n\to\infty} \frac{\lambda_n}{h^{-1}(\log n)} = 1. \qquad \square$$

REMARK 2.8. If f above is asymptotically like the inverse of a fractional polynomial, that is, if there is a $d > 3$ such that $\lim_{x\to\infty} f(x)x^d \in (0, \infty)$, then easy computations yield $p(\lambda, 0) \sim c\lambda^{2-d}$, implying that $\lambda_n \sim \sqrt[d]{cn}$, for some constants $c > 0$.

REMARK 2.9. The additional summand σ^2 in the benchmark $B_n(\theta, \sigma^2)$ can be moderately increased without changing the asymptotics of the thresholds. Indeed, if σ^2 is replaced by $c_n\sigma^2$, with $c_n = O((\log n)^\beta)$, $\beta > 0$, then after changing q, above, to $q(\lambda) := (\lambda^2 + \sigma^2)/p(\lambda, 0) - c_n$, the optimal thresholds are still the solution of $q(\lambda) = n$, and the asymptotic behavior of q^{-1} is not changed. Thus the asymptotic behavior of the thresholds is not changed either.

Although there are in general no closed form formulas for the thresholds in Theorem 2.1, it is quite easy to compute numerical approximations. The results of such computations, using Mathematica, are presented in Figure 1 (see also [13] for the normal case). The types of noise distributions considered are the normal distribution, the Laplace distribution and the distribution with the density $c_1 \exp(-c_2\sqrt{|x|})$. Additionally, the optimal thresholds for densities with polynomial decay, the uniform distribution and some mixtures of them are also given. All distributions are scaled to have variance 1. In Figure 1 the densities are only labeled by their functional part, that is, the actual density is the functional part

$n \to$	32	128	512	2048	65536	2^{24}	2^{32}
ϕ	1.28	1.67	2.04	2.40	3.22	4.35	5.31
$\exp(-\lvert x\rvert)$	1.58	2.19	2.85	3.55	5.43	8.70	12.15
$\exp(-\sqrt{\lvert x\rvert})$	2.18	3.26	4.60	6.21	11.6	24.2	42.1
$\mathbf{1}_{-1,1}(x)$	1.04	1.26	1.42	1.53	1.66	1.72	1.74
$1/(x^{10}+1)$	1.11	1.40	1.68	1.99	2.96	5.53	10.3
$1/(x^4+1)$	1.99	3.28	5.30	8.46	27.0	171	1088
$1/((x+20)^4+1)$	2.78	4.67	7.64	12.3	40.0	256	1625
$1/((x+0.1)^4+1)$	2.08	3.44	5.57	8.91	28.5	181	1149
$999\phi(x) + 1/(x+1)^4$	1.28	1.67	2.05	2.42	3.51	22.0	141
$99\phi(x) + 1/(x+1)^4$	1.29	1.69	2.10	2.57	7.23	47.9	306
$9\phi(x) + 1/(x+1)^4$	1.37	1.93	2.89	4.82	16.2	105	666

FIG. 1. *The optimal thresholds for some densities.*

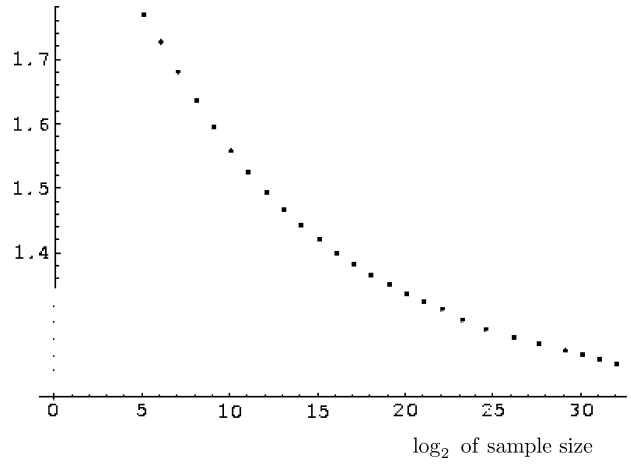

FIG. 2. *Ratio of asymptotic threshold and optimal threshold for the normal distribution.*

scaled such that it has variance 1. For example, $\exp(-x^2)$ is the functional part of the standard normal density ϕ. The rationale for a maximal $n = 2^{32}$ is that most of today's (December 1999) computers are not able to work with datasets larger than 2^{32} (32 bit address bus).

Figure 2 (resp. 3) shows the ratio of the asymptotic threshold and of the optimal threshold for the normal distribution (resp., the Laplace distribution). The horizontal axis represents the base 2 logarithm of the sample size. As one sees, the asymptotics in Theorem 2.7 work very slowly, since for the normal distribution (resp. the Laplace distribution) and 2^{23} (resp. 2^{27}) data points, this ratio is approximately 1.3.

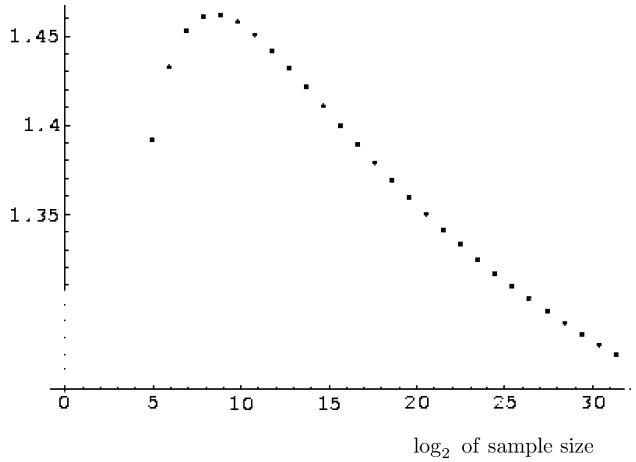

FIG. 3. *Ratio of asymptotic threshold and optimal threshold for the Laplace distribution.*

3. Near minimaxity. If the noise is normal, thresholding achieves the minimax rate in the class of all estimators, that is,

$$\lim_{n\to\infty} \inf_{\widehat{\theta}} \sup_{\theta\in\mathbb{R}^n} \frac{E\|\widehat{\theta}-\theta\|^2}{B_n(\theta,\sigma^2)} \frac{1}{\Lambda_n} = 1,$$

where the infimum is taken over all estimators and where Λ_n is now computed for the normal distribution [13]. For a special class of distributions one can also show that soft thresholding is asymptotically "near" minimax; that is, the 1 on the above right-hand side is replaced by a constant. This result is a natural consequence of Theorem 2.7 and of our next result (again the normal case is recovered by taking below $h(x) = x^2/2$).

THEOREM 3.1. *Let $Y_i = \theta_i + z_i$, $i = 1, \dots, n$, where the z_i are i.i.d. random variables with $Ez_1 = 0$ and $Ez_1^2 = \sigma^2$ and law μ. Let μ be absolutely continuous with density f of the form $f(x) = C\exp(-h(x))$ (C a normalizing constant), where h is even, continuous and increasing on $[0,\infty)$. Further, let*

$$\limsup_{x\to\infty} \frac{h^{-1}(x)}{h^{-1}(x - 2\log x)} = 1.$$

Then

$$\liminf_{n\to\infty} \inf_{\widehat{\theta}} \sup_{\theta\in\mathbb{R}^n} \frac{E\|\widehat{\theta}-\theta\|^2}{B_n(\theta,\sigma^2)} \frac{\sigma^2}{(h^{-1}(\log n))^2} \geq 1,$$

where the infimum is taken over all estimators $\widehat{\theta}$ of θ.

PROOF. We prove this bound by computing Bayes risks (a strategy already used in the proof of Theorem 3 in [13]). For $0 < \varepsilon < 1$ and $a > 0$, let $F_{\varepsilon,a} := \varepsilon\delta_a + (1-\varepsilon)\delta_0$, where δ_c denotes the Dirac measure with unit mass at c. The a priori measure for $\theta \in \mathbb{R}^n$ is $Q_n := \bigotimes_{i=1}^n F_{\varepsilon_n,a_n}$, with ε_n and a_n specified later. For now, it suffices to assume that $\varepsilon_n \to 0$ and $a_n \to \infty$. First we consider the one-dimensional case, and compute the Bayes risk for estimating $\theta_1 \in \mathbb{R}$ given $Y_1 = \theta_1 + z_1$, where the a priori measure for θ_1 is $F_{\varepsilon,a}$. Let $M := F_{\varepsilon,a} * f$; that is, for any A, B, Borel sets in $\mathbb{R}$,

$$M(A,B) = (1-\varepsilon)\delta_0(A)\int_B f(x)\,dx + \varepsilon\delta_a(A)\int_B f(x-a)\,dx.$$

Let Π_1 and Π_2 be the projection from $\mathbb{R}^2$ to the first, respectively, the second coordinate. In this context, the Bayes estimator for θ is

$$\begin{aligned} d_{\varepsilon,a}(x) = E_M(\Pi_1|\Pi_2 = x) &= \frac{0(1-\varepsilon)f(x) + a\varepsilon f(x-a)}{(1-\varepsilon)f(x) + \varepsilon f(x-a)} \\ &= \frac{\varepsilon f(x-a)}{\varepsilon f(x-a) + (1-\varepsilon)f(x)} a. \end{aligned} \tag{3.1}$$

Thus,

$$
\begin{aligned}
&E_{F_{\varepsilon,a}} E_{\theta_1} (d_{\varepsilon,a} - \theta_1)^2 \\
&\quad = \varepsilon \int_{-\infty}^{+\infty} \big(d_{\varepsilon,a}(x) - a\big)^2 f(x-a)\,dx + (1-\varepsilon) \int_{-\infty}^{+\infty} d_{\varepsilon,a}(x)^2 f(x)\,dx \\
&\quad \geq \varepsilon a^2 \int_{-\infty}^{+\infty} \left(1 - \frac{\varepsilon f(x-a)}{\varepsilon f(x-a) + (1-\varepsilon) f(x)}\right)^2 f(x-a)\,dx \qquad (3.2) \\
&\quad = \varepsilon a^2 \int_{-\infty}^{+\infty} \left(\frac{(1-\varepsilon) f(x)}{\varepsilon f(x-a) + (1-\varepsilon) f(x)}\right)^2 f(x-a)\,dx \\
&\quad = (1-\varepsilon)^2 \varepsilon a^2 \int_{-\infty}^{+\infty} \frac{f(x)^2}{(\varepsilon f(x-a) + (1-\varepsilon) f(x))^2} f(x-a)\,dx.
\end{aligned}
$$

Let us further lower bound (3.2). For any $\alpha \in (0,1)$, there exists a $c > 0$ such that $\int_{-c}^{c} f(x)\,dx \geq \alpha$, while for any $\beta > 0$, assume that there exist (as will be shown below) $\varepsilon > 0$ and $a > 0$ such that $\beta f(a+c) \geq \frac{\varepsilon}{1-\varepsilon} f(0)$. Then, since f is even and decreasing on $[0, +\infty)$, $\beta f(x) \geq \frac{\varepsilon}{1-\varepsilon} f(x-a)$, for all $x \in (a-c, a+c)$. In turn, this implies that for all $x \in (a-c, a+c)$,

$$
\frac{f(x)^2}{(\varepsilon f(x-a) + (1-\varepsilon) f(x))^2} \geq \frac{f(x)^2}{((1-\varepsilon) f(x)(1+\beta))^2},
$$

and using (3.2),

$$
E_{F_{\varepsilon,a}} E_{\theta_1} (d_{\varepsilon,a} - \theta_1)^2 \geq \frac{(1-\varepsilon)^2}{(1-\varepsilon)^2} \frac{\alpha}{(1+\beta)^2} \varepsilon a^2 = \frac{\alpha}{(1+\beta)^2} \varepsilon a^2. \qquad (3.3)
$$

Let us now show how to apply the above inequalities to the multivariate Bayes case. Let α (hence c) and β be fixed, let ε_n, a_n be sequences such that $n\varepsilon_n \to \infty$ and $\beta f(a_n + c) \geq \frac{\varepsilon_n}{1-\varepsilon_n} f(0)$, and finally let $m_n := (n\varepsilon_n)^{2/3}$, $N_n := \#\{\theta_i \neq 0, i = 1, \ldots, n\}$, $A_n := \{N_n \leq n\varepsilon_n + m_n\}$, and $p_n := Q_n(A_n^c)$, where # denotes cardinality. We now prove that $p_n = Q_n(N_n - n\varepsilon_n > m_n) = o(\varepsilon_n)$ by using Bennett's inequality [35], page 851. This inequality provides the first step below, where the function $k(x) := 2((1+x)\log(1+x) - x)/x^2$ is continuous, decreasing on $(0, \infty)$ and such that $k(0) = 1$,

$$
\begin{aligned}
p_n &\leq \exp\left(-\frac{m_n/\sqrt{n}}{2\varepsilon_n(1-\varepsilon_n)} k\left(\frac{m_n/\sqrt{n}}{\varepsilon_n(1-\varepsilon_n)\sqrt{n}}\right)\right) \\
&\leq \exp\left(-\frac{(\varepsilon_n n)^{2/3}}{2\varepsilon_n\sqrt{n}} k\left(\frac{(\varepsilon_n n)^{2/3}}{n\varepsilon_n(1-\varepsilon_n)}\right)\right) \\
&= \exp\left(-\frac{1}{2}\varepsilon_n^{-1/3} n^{1/6} k\left(\frac{(\varepsilon_n n)^{-1/3}}{1-\varepsilon_n}\right)\right) \qquad (3.4)
\end{aligned}
$$

$$\leq \exp\left(-\frac{1}{4}\varepsilon_n^{-1/3}n^{1/6}\right)$$

[for n large enough since $n\varepsilon_n \to \infty$ and $k(0) = 1$]

$$= o(\varepsilon_n).$$

We are now ready to tackle the Bayes risk of Q_n. Let d_n be the ($\mathbb{R}^n$-valued) Bayes estimator for Q_n and let d_n^i, $i = 1, \ldots, n$, be its ith coefficient. In view of (3.1) it is clear that $0 \leq d_n^i \leq a_n$. Next,

$$\begin{aligned}
E_{Q_n}E_\theta \frac{\|d_n(Y)-\theta\|^2}{B_n(\theta,\sigma^2)} &\geq \frac{1}{\sigma^2(1+n\varepsilon_n+m_n)}\, E_{Q_n}E_\theta \sum_{i=1}^n \big(d_n^i(Y)-\theta_i\big)^2 \mathbf{1}_{A_n} \\
&\geq \frac{1}{\sigma^2(1+n\varepsilon_n+m_n)}\left(E_{Q_n}E_\theta \sum_{i=1}^n \big(d_n^i(Y)-\theta_i\big)^2 - np_na_n^2\right) \\
&\qquad\qquad [\text{since } E_\theta(d_n^i(Y)-\theta_i)^2\mathbf{1}_{A_n^c} \leq p_na_n^2] \\
&= \frac{1}{\sigma^2(1+n\varepsilon_n+m_n)}\left(\sum_{i=1}^n E_{Q_n}E_{\theta_i}\big(d_n^i(Y)-\theta_i\big)^2 - np_na_n^2\right) \\
&\geq \frac{1}{\sigma^2(1+n\varepsilon_n+m_n)}\left(n\varepsilon_n a_n^2\frac{\alpha}{(1+\beta)^2} - np_na_n^2\right) \\
&\sim \frac{\alpha}{(1+\beta)^2\sigma^2}a_n^2,
\end{aligned} \tag{3.5}$$

using (3.3) and since $p_na_n^2 = o(\varepsilon_n a_n^2)$ and $m_n = o(n\varepsilon_n)$. Let us now choose ε_n and a_n. Remember that ε_n has to satisfy $n\varepsilon_n \to \infty$ and that a_n must be such that $f(a_n + c) \geq \frac{\varepsilon_n}{\beta(1-\varepsilon_n)} f(0)$. Thus, if $n\varepsilon_n = \log n$ and since f^{-1} is decreasing, we can optimally choose

$$a_n = h^{-1}\left(\log n - \log\log n + \log\left(1 - \frac{\log n}{n}\right) + \log\beta - \log f(0)\right) - c. \tag{3.6}$$

Because of the conditions imposed on h^{-1}, we have $a_n \sim h^{-1}(\log n)$. Since α and β are arbitrary, the theorem follows by letting $\alpha \to 1$ and $\beta \to 0$. □

REMARK 3.2. Above, the time domain i.i.d. assumption which is natural in view of the regression model (1.2) could be replaced by independence and the (less natural) requirement that the ratio of two arbitrary noise densities is uniformly bounded above and below. It should also be noted that the proof just presented readily adapts to a general symmetric, continuous and decreasing noise density f, since a lower bound for the Bayes risk similar to (3.5) can be obtained where now $a_n = f^{-1}(\varepsilon_n/(\beta(1-\varepsilon_n))f(0)) - c$. Above, the condition $\limsup_{x\to\infty} h^{-1}(x)/(h^{-1}(x - 2\log x)) = 1$ (where actually $\limsup = \lim$, since

h^{-1} is increasing on $[0, +\infty))$ was mainly imposed to simplify the resulting asymptotic behavior of a_n in (3.6) and is satisfied if h grows at most as fast as a fractional polynomial. Combining the last two theorems we see that if the noise in the data is i.i.d. and its density is asymptotically like $\exp(-h(x))$ where h is a fractional polynomial, then soft thresholding is optimal in the minimax sense, it has the asymptotically best ratio of risk and benchmark.

REMARK 3.3. Again, and as previously noticed, the benchmark can be moderately increased. One of the last steps in the proof of the previous theorem asserts that

$$\frac{1}{\sigma^2(1+n\varepsilon_n+m_n)}\left(n\varepsilon_n a_n^2\frac{\alpha}{(1+\beta)^2}-np_n a_n^2\right)\sim\frac{\alpha}{(1+\beta)^2\sigma^2}a_n^2.$$

The constant 1 in the denominator on the left-hand side can be replaced by c_n. The $\sim$ remains valid if $c_n/(\varepsilon_n n)\to 0$. To keep the same rate for a_n it is necessary that $\log\varepsilon_n\sim-\log n$. Thus, if for example $c_n\sim(\log n)^p$, then with $\varepsilon_n=(\log n)^{p+1}/n$ we get the same asymptotic rate for a_n and for the Bayes risk for Q_n.

In contrast to Theorem 3.1, our next result [which applies to the wavelet domain model (1.3)] is only about the rate of the ratios. The idea of the proof is the same in both results, but the proof is now complicated by the fact that we do not have a closed form expression for the density of the noise in the empirical wavelet coefficients. Since we have to rely on rough bounds for these tails, we only get a statement about the rate.

THEOREM 3.4. *Let the observations $X_i=f_i+e_i$, $i=0,\dots,n-1=2^m-1$, be given, where the f_i are parameters of interest and the e_i are i.i.d. random variables with law μ and variance σ^2. Let μ be absolutely continuous with density g of the form $g(x)=C\exp(-h(x))$ (C a normalizing constant), where h is even, continuous and increasing on $[0,\infty)$. Further, let $\limsup_{x\to\infty}h(cx)/h(x)<\infty$, for all $c>0$. Apply a periodic wavelet transform W_n to these observations, taking for every n the same compactly supported generating wavelet. Let $\theta=W_n(f)$. Then*

$$\liminf_{n\to\infty}\inf_{\widehat{\theta}}\sup_{\theta\in\mathbb{R}^n}\frac{E\|\widehat{\theta}-\theta\|^2}{B_n(\theta,\sigma^2)}\frac{1}{(h^{-1}(\log n))^2}>0,$$

where the infimum is taken over all estimators $\widehat{\theta}$ of θ.

PROOF. We apply a wavelet transform W_n derived from a multiresolution analysis generated by a common compactly supported wavelet. The fixed filter length is N, N is even, and $Y_{j,k}$, $\theta_{j,k}$ and $z_{j,k}$, $0\le k\le 2^j-1$, $0\le j\le m-1$, are

respectively the kth wavelet coefficient at the level j, of $W_n(X)$, $W_n(f)$ and $W_n(e)$. Let $\widehat{\theta}_{j,k}$ be an estimator for $\theta_{j,k}$, where $\widehat{\theta}_{j,k}$ may depend on all the X_i. We compute below an asymptotic lower bound for $\sup_\theta(\sum_{j,k} E|\widehat{\theta}_{j,k} - \theta_{j,k}|^2)/B_n(\theta, \sigma^2)$, by computing a Bayes risk. For $0 < \varepsilon < 1$ and $a > 0$, let $F_{\varepsilon,a}$ be defined as in the proof of the previous theorem. The a priori measure for $\theta_{j,k}$ is F_{ε_n,a_n} if $j = m-1$, and δ_0 otherwise. For now, it suffices to know that $\varepsilon_n \to 0$ and that $a_n \to \infty$. The a priori measure Q_n for $\theta \in \mathbb{R}^n$ is the product measure of the a priori measures of the coordinates. In the sequel, we omit the level coefficient $m-1$, since we are only concerned with the coefficients in this level. Let us now observe the following: For n greater than a certain bound, the filter coefficients $d_0, \ldots, d_{N-1}$ used to compute the $\theta_{j,k}$ no longer depend on n. Thus $\theta_i = \sum_{\ell=0}^{N-1} d_\ell f_{2i+\ell}$ where the indices for f are considered modulo n, since we are using a periodic wavelet transform. On the other hand, given the a priori measure Q_n (that is, assuming that the wavelet coefficients for the levels $j < m-1$ are 0) one computes the f_i by

$$f_i := \sum_{\ell=0,\ \ell \text{ even}}^{N-1} d_\ell \theta_{(i-\ell)/2} \qquad \text{if } i \text{ is even,}$$

and

$$f_i := \sum_{\ell=0,\ \ell \text{ odd}}^{N-1} d_\ell \theta_{(i-\ell)/2} \qquad \text{if } i \text{ is odd.}$$

This follows immediately from the fact that W_n is orthonormal, that is, $W_n^T = W_n^{-1}$. For $0 \le i \le n/2 - 1$, set

$$Q_n^i(\cdot) := \frac{Q_n(\cdot \cap \{\theta_{i+j} = 0, 1 \le |j| \le N/2 - 1\})}{Q_n(\{\theta_{i+j} = 0, 1 \le |j| \le N/2 - 1\})},$$

where $0/0$ is understood as 0. If Q_n^i is the a priori measure, then θ_i and f_j are independent whenever $j < 2i$ or $j > 2i + N - 1$. Hence the Bayes estimator for θ_i given the a priori measure Q_n^i only depends on $f_{2i}, \ldots, f_{2i+N-1}$. As one easily checks the projection of Q_n^i on $f_{2i}, \ldots, f_{2i+N-1}$ is $\varepsilon_n \delta_{(d_0,\ldots,d_{N-1})a_n} + (1-\varepsilon_n)\delta_{(0,\ldots,0)}$. Hence the Bayes estimator b_i for θ_i given Q_n^i is

$$b_i(x) := \frac{g_{a_n}(x)\varepsilon_n}{g_{a_n}(x)\varepsilon_n + (1-\varepsilon_n)g_0(x)} a_n, \qquad x \in \mathbb{R}^N,$$

where g_{a_n} (resp. g_0) is the density of $X_{2i}, \ldots, X_{2i+N-1}$ if $(f_{2i}, \ldots, f_{2i+N-1}) = a_n(d_0, \ldots, d_{N-1})$ [resp. if $(f_{2i}, \ldots, f_{2i+N-1}) = (0, \ldots, 0)$]. In other words, $g_{a_n}(x) = \prod_{\ell=0}^{N-1} g(x_\ell - a_n d_\ell)$ and $g_0(x) = \prod_{\ell=0}^{N-1} g(x_\ell)$. Now let $\widehat{\theta}_i$ be an estimator for θ_i. Then

$$E_{Q_n} E_{\theta_i}(\widehat{\theta}_i - \theta_i)^2 \ge (1-\varepsilon_n)^{N-2} E_{Q_n^i} E_{\theta_i}(b_i - \theta_i)^2. \tag{3.7}$$

Next, we proceed to compute the Bayes risk of b_i given the a priori measure Q_n^i:

$$
\begin{aligned}
E_{Q_n^i} E_{\theta_i}(b_i - \theta_i)^2 &= (1-\varepsilon_n) E_{\theta_i=0} b_i^2 + \varepsilon_n E_{\theta_i = a_n}(b_i - a_n)^2 \\
&\geq \varepsilon_n \int_{\mathbb{R}^N} \left(\frac{g_{a_n}(x)\varepsilon_n}{g_0(x)(1-\varepsilon_n) + g_{a_n}(x)\varepsilon_n} a_n - a_n \right)^2 g_{a_n}(x)\,dx \qquad (3.8) \\
&= \varepsilon_n a_n^2 \int_{\mathbb{R}^N} \left(\frac{g_0(x)(1-\varepsilon_n)}{g_0(x)(1-\varepsilon_n) + g_{a_n}(x)\varepsilon_n} \right)^2 g_{a_n}(x)\,dx.
\end{aligned}
$$

Set $d_{\max} := \max_{i=0,\ldots,N-1} |d_i|$. Let $0 < \alpha < 1$. Then there exists a $c > 0$ (depending on α) such that $\int_{[-c,c]^N} g_0(x)\,dx \geq \alpha$. Now define $I_n := [-c,c]^N + a_n(d_0, \ldots, d_{N-1})$, and assume that ε_n and a_n are such that $(1-\varepsilon_n) g_0(x) \geq 2\varepsilon_n g_{a_n}(x)$, for all $x \in I_n$. Thus using (3.8),

$$
E_{Q_n^i} E_{\theta_i}(b_i - \theta_i)^2 \geq \varepsilon_n a_n^2 \int_{I_n} \left(\frac{g_0(x)(1-\varepsilon_n)}{g_0(x)(1-\varepsilon_n)3/2} \right)^2 g_{a_n}(x)\,dx \geq \frac{4}{9}\varepsilon_n a_n^2 \alpha,
$$

which implies via (3.7)

$$
E_{Q_n} E_{\theta_i}(\widehat{\theta}_i - \theta_i)^2 \geq \tfrac{4}{9}(1-\varepsilon_n)^{N-2} \varepsilon_n a_n^2 \alpha.
$$

Hence for α large enough and n greater than a certain bound, the Bayes risk for estimating a single θ_i is greater than or equal to $\frac{4}{10}\varepsilon_n a_n^2$. Let ε_n and a_n be sequences such that $(1-\varepsilon_n) g_0(x) \geq 2\varepsilon_n g_{a_n}(x)$, for all $x \in I_n$, with also $\lim_{n\to\infty} n\varepsilon_n = +\infty$. Set $m_n := (n\varepsilon_n/2)^{2/3}$, and as in the previous proof let $N_n := \#\{\theta_i \neq 0,\ i = 0, \ldots, n/2 - 1\}$, $A_n := \{N_n \leq n\varepsilon_n/2 + m_n\}$, and $p_n := Q_n(A_n^c)$. Recall also from (3.4) that $p_n = o(\varepsilon_n)$. Let $\widehat{\widehat{\theta}}_{j,k}$ be an estimator for $\theta_{j,k}$ and, without loss of generality, assume again that $0 \leq \widehat{\theta}_{m-1,k} \leq a_n$. Now, given Q_n, that is, assuming that the wavelet coefficients for the levels $j < m-1$ are 0, it follows that

$$
\begin{aligned}
&E_{Q_n} E_\theta \frac{\sum_{j,k} |\widehat{\theta}_{j,k} - \theta_{j,k}|^2}{B_n(\theta, \sigma^2)} \\
&\quad \geq E_{Q_n} E_\theta \frac{\sum_k |\widehat{\theta}_k - \theta_k|^2}{\sigma^2(1+N_n)} \\
&\qquad\qquad \text{(leaving the index } m-1 \text{ out, that is, } \theta_k := \theta_{m-1,k}) \\
&\quad \geq \frac{1}{\sigma^2(1+n\varepsilon_n/2+m_n)} E_{Q_n} E_\theta \sum_k |\widehat{\theta}_k - \theta_k|^2 \mathbf{1}_{A_n} \\
&\quad \geq \frac{1}{\sigma^2(1+n\varepsilon_n/2+m_n)} E_{Q_n} E_\theta \sum_k |\widehat{\theta}_k - \theta_k|^2 - \frac{n}{2} p_n a_n^2 \\
&\qquad\qquad [\text{since } E_\theta \mathbf{1}_{A_n^c} |\widehat{\theta}_k - \theta_k|^2 \leq Q_n(A_n^c) a_n^2 = p_n a_n^2] \\
&\quad = \frac{1}{\sigma^2(1+n\varepsilon_n/2+m_n)} \sum_k E_{Q_n} E_\theta |\widehat{\theta}_k - \theta_k|^2 - \frac{n}{2} p_n a_n^2
\end{aligned}
$$

$$\geq \frac{1}{\sigma^2(1+n\varepsilon_n/2+m_n)}\left(\frac{n}{2}\frac{4}{10}\varepsilon_n a_n^2 - \frac{n}{2}p_n a_n^2\right)$$

(for n large enough)

$$\geq \frac{1}{3\sigma^2}\,a_n^2 \qquad \text{[for } n \text{ large enough and since } p_n = o(\varepsilon_n)\text{]}.$$

Take $\varepsilon_n := \frac{\log n}{n}$, assume n is so large that $\varepsilon_n < 1/2$, and choose

$$a_n := h^{-1}\left(\frac{1}{N}(\log n - \log\log n) - \log\bigl(g(0)\sqrt[N]{4}\bigr)\right) - c.$$

It then follows that

$$h(a_n + c) = -\log\bigl(g(0)\sqrt[N]{4}\bigr) + \frac{1}{N}(\log n - \log\log n),$$

equivalently, $\exp(-h(a_n + c)) = g(0)\sqrt[N]{\log n/n}\,\sqrt[N]{4}$, that is, $g(a_n + c) = g(0)\sqrt[N]{\varepsilon_n}\,\sqrt[N]{4}$. Since $\varepsilon_n < 1/2$ and $|d_{\max}| \leq 1$,

$$g(d_{\max}\,a_n + c) \geq g(0)\sqrt[N]{\varepsilon_n}\,\sqrt[N]{2/(1-\varepsilon_n)};$$

thus

$$(1-\varepsilon_n)\prod_{\ell=0}^{N-1} g(d_\ell a_n + c) \geq 2\varepsilon_n g(0)^N,$$

and finally

$$(1-\varepsilon_n)g_0(x) \geq 2\varepsilon_n g_{a_n}(x)$$

for all $x \in I_n$. Our choice of a_n has fulfilled the required conditions, and since $a_n \sim c_1 h^{-1}(\log n)$ for some constant $c_1 > 0$, the theorem is proved. □

4. Compactly supported distributions. In the previous section, we computed lower asymptotic bounds for thresholding for some classes of noise. To do so, we made use of the facts that half of the coefficients are in the finest level, and that the distribution of the noise in this level is "close" to the distribution of the original noise. We consider now compactly supported noise and show that the asymptotic performance of the wavelet domain thresholding is not better than for normal noise. Here we will use the fact that the distribution of the noise in the $\sqrt{n}$ coarser level coefficients is "close" to the normal distribution. From the discussion before Proposition 2.6, the result below can be complemented by the fact that for compactly supported noise, Λ_n (as defined below) is such that $\Lambda_n = O(\log n)$ (see also the next section).

THEOREM 4.1. *Let the observations $X_i = f_i + e_i$, $i = 1, \dots, n = 2^m$, be given, where the f_i are parameters of interest and the e_i are zero mean, compactly supported (on $[-M, M]$) i.i.d. random variables with variance σ^2. Apply a periodic wavelet transform W_n to these observations, taking for every n the same compactly supported generating wavelet. Assume moreover that the corresponding wavelet basis is Hölder continuous of index $\beta > 0$. Let $Y = W_n(X)$, let $\theta = W_n(f)$ and let*

$$\Lambda_n := \inf_{(\lambda_{\cdot,\cdot}) \in \mathbb{R}^n} \sup_{\theta \in \mathbb{R}^n} \frac{\sum_{j,k} E|T^S_{\lambda_{j,k}}(Y_{j,k}) - \theta_{j,k}|^2}{B_n(\theta, \sigma^2)}. \tag{4.1}$$

Then $\liminf_{n\to\infty} \Lambda_n/(2\log n) \geq 1$.

PROOF. Let $z = W_n(e)$, where the wavelet transform W_n is derived from a multiresolution analysis generated by a common compactly supported wavelet ψ, and the fixed filter length is N. To prove the result, we need to lower bound $p(\lambda, 0)$ corresponding to the noise $(z_{j,k})$ in the wavelet coefficients, where again $z_{j,k} = \sum_{i=1}^n w^{(n)}_{j,k,i} e_i$, with $\sum_{i=1}^n (w^{(n)}_{j,k,i})^2 = 1$. We first show that $\max_{k,i,j\leq h} |w^{(n)}_{j,k,i}| \leq C 2^{(j-m)/2}$, for some constant C and where for a fixed $q \in (0,1)$, $h = h(n) = [\log_2(n^q)]$ ($[\cdot]$ denotes integer part). This first step is needed to apply known estimates to control the tail behavior of $z_{j,k}$. Recall that if W_n is generated by a compactly supported wavelet not necessarily adapted to an interval, then $w^{(n)}_{j,k,i} = \langle \psi_{j,k}, \varphi_{m,i} \rangle$, where φ is the scaling function associated to the wavelet ψ, and again $n = 2^m$. It is also well known that if φ is Hölder continuous with exponent β, then

$$\sup_{i,k} |2^{(m-j)/2} \langle \psi_{j,k}, \varphi_{m,i} \rangle - \psi(2^{j-m} i - k)| \leq C_1 2^{\beta(j-m)}, \tag{4.2}$$

for some constant C_1 and $m - j > j_0$, for some j_0 ([8], page 205). But W_n is a periodic wavelet transform adapted to $[0, 1]$; that is, for $j \geq 0$ and $0 \leq k < 2^j$, $\psi_{j,k}$ and $\varphi_{j,k}$ are replaced by

$$\psi^{\mathrm{per}}_{j,k}(x) := \sum_{i\in\mathbb{Z}} \psi_{j,k}(x + i) \quad \text{and} \quad \varphi^{\mathrm{per}}_{j,k}(x) := \sum_{i\in\mathbb{Z}} \varphi_{j,k}(x + i);$$

thus $w^{(n)}_{j,k,i} = \langle \psi^{\mathrm{per}}_{j,k}, \varphi^{\mathrm{per}}_{m,i} \rangle$. Since for m large enough, $\varphi^{\mathrm{per}}_{m,i}(x) = \varphi_{m,i}(x - [x])$ and since in the construction of each $\psi^{\mathrm{per}}_{\cdot,\cdot}$ only at most N wavelets are involved, a bound similar to (4.2) continues to hold, but with C_1 replaced by a constant C_2 depending on N and C_1. Hence

$$\max_{k,i} |w^{(n)}_{j,k,i}| \leq 2^{(j-m)/2} N \|\psi\|_\infty + C_2 2^{(j-m)(\beta+1/2)} \leq C_3 2^{(j-m)/2}, \tag{4.3}$$

where $C_3 := C_2 + N\|\psi\|_\infty$. Now

$$\sup_{\theta\in\mathbb{R}^n} \frac{\sum_{j,k} E|T^S_{\lambda_{j,k}}(Y_{j,k}) - \theta_{j,k}|^2}{B_n(\sigma^2,\theta)} \geq \max\left(\frac{\sum_{j\leq h,k} E(T^S_{\lambda_{j,k}}(z_{j,k}))^2}{\sigma^2}, \frac{\sum_{j\leq h,k}(\lambda^2_{j,k}+\sigma^2)}{(2^{h+1}+1)\sigma^2}\right), \tag{4.4}$$

where this lower bound is obtained by choosing, respectively, $\theta_{j,k} = 0$ for all j,k and

$$\theta_{j,k} = \begin{cases} \infty, & j \leq h, \\ 0, & \text{elsewhere.} \end{cases}$$

Because of (4.4), from now on we will only be concerned with the levels $0,\dots,h$, and in the rest of the proof when (j,k) is an index then implicitly $j \leq h$.

To bound the first term in the maximum in (4.4) we use the following version of Kolmogorov's converse exponential inequality (see [36], Section 5.2).

Let (X_i) be a finite sequence of independent, zero mean random variables such that $\|X_i\|_\infty \leq D$, for all i. Then for every $\gamma > 0$, there exist positive reals $K(\gamma)$ (large enough) and $\varepsilon(\gamma)$ (small enough) such that for every t satisfying $t \geq K(\gamma)b$ and $tD \leq \varepsilon(\gamma)b^2$,

$$P\left(\sum_i X_i > t\right) \geq \exp\left(-\frac{(1+\gamma)t^2}{2b^2}\right), \tag{4.5}$$

where $b^2 = \sum_i EX_i^2$.

Let us now show how to use (4.5), and let $\gamma > 0$. Recall that $\|e_k\|_\infty \leq M < \infty$ for all k, that by (4.3) $\max_{j\leq h,k,i} \|w^{(n)}_{j,k,i}e_i\|_\infty \leq C_3 M 2^{(h-m)/2}$, and that $\sum_i E(w^{(n)}_{j,k,i}e_i)^2 = \sigma^2$. Thus for $t \geq K(\gamma)\sigma$ and $t \leq \varepsilon(\gamma)\sigma^2 2^{(m-h)/2}/(C_3M)$,

$$P(|z_{j,k}| > t) \geq 2\exp\left(-\frac{(1+\gamma)t^2}{2\sigma^2}\right). \tag{4.6}$$

In addition to (4.6), let us also state two estimates we need. First, by Chebychev's inequality,

$$P(|z_{j,k}| \geq \lambda_{j,k} + 1) \leq E\big(T^S_{\lambda_{j,k}}(z_{j,k})\big)^2. \tag{4.7}$$

Second, thanks to the uniform convergence in the CLT with rate (see [34], Section 5.2) for all $x > 0$, there exists an $n_0 = n_0(x)$ such that for all $n \geq n_0$,

$$2P(|z_{j,k}| \geq \sigma x) \geq 1 - \Phi(x), \tag{4.8}$$

where Φ is the standard normal distribution function.

Now let $(\lambda^*_{j,k,n})$ be a set of optimal thresholds for the right-hand side of (4.1) and let $0 < \alpha < 1$. Recall that $j \leq h$, hence $\#\{\lambda_{j,k,n}, j \leq h\} = 2^{h+1}$ (again,

# denotes cardinality) and thus at least one of the following three conditions holds:

$$\#\left\{\lambda^*_{j,k,n} \in \left(K(\gamma)\sigma, \frac{\varepsilon(\gamma)\sigma^2 2^{(m-h)/2}}{C_3 M}\right)\right\} \geq (1-\alpha)2^{h+1}, \tag{4.9}$$

or

$$\#\{\lambda^*_{j,k,n} \leq K(\gamma)\sigma\} \geq \alpha 2^h, \tag{4.10}$$

or

$$\#\left\{\lambda^*_{j,k,n} \geq \frac{\varepsilon(\gamma)\sigma^2 2^{(m-h)/2}}{C_3 M}\right\} \geq \alpha 2^h. \tag{4.11}$$

Combining (4.4) and (4.7) leads to

$$\Lambda_n \geq \max\left(\frac{\sum_{j\leq h,k} P(|z_{j,k}| \geq \lambda^*_{j,k,n}+1)}{\sigma^2}, \frac{\sum_{j,k}((\lambda^*_{j,k,n})^2+\sigma^2)}{(2^{h+1}+1)\sigma^2}\right). \tag{4.12}$$

Assuming (4.9) and using (4.6), (4.12) becomes

$$\Lambda_n \geq \max\left(\sum_{j\leq h,k}\frac{2}{\sigma^2}\exp\left(-\frac{(1+\gamma)(\lambda^*_{j,k,n}+1)^2}{2\sigma^2}\right), \frac{\sum_{j\leq h,k}((\lambda^*_{j,k,n})^2+\sigma^2)}{(2^{h+1}+1)\sigma^2}\right)$$

$$\geq \inf_{(\lambda_{\cdot,\cdot,\cdot})\in\mathbb{R}^n} \max\left(\sum_{j\leq h,k}\frac{2}{\sigma^2}\exp\left(-\frac{(1+\gamma)(\lambda_{j,k,n}+1)^2}{2\sigma^2}\right), \frac{\sum_{j\leq h,k}((\lambda_{j,k,n})^2+\sigma^2)}{(2^{h+1}+1)\sigma^2}\right), \tag{4.13}$$

where the sums are over the $\lambda^*_{j,k,n}$ and $\lambda_{j,k,n}$ satisfying (4.9).

It is easy to see that the minimum on the right-hand side of (4.13) is achieved if all the $\lambda_{j,k,n}$ are the same and equal to the solution λ^*_n of

$$\frac{2}{\sigma^2}\exp\left(-\frac{(1+\gamma)(\lambda_n+1)^2}{2\sigma^2}\right) = \frac{(\lambda_n^2+\sigma^2)}{(2^{h+1}+1)\sigma^2}.$$

Now computations as in the proof of Theorem 2.7 show that λ^*_n behaves asymptotically like

$$\sqrt{\frac{2\sigma^2\log 2^{h+1}}{(1+\gamma)}} \sim \sqrt{\frac{2\sigma^2 q\log n}{(1+\gamma)}},$$

where the asymptotic $\sim$ holds by our choice of $h = h(n)$. Hence (4.9) and (4.13) lead to

$$\Lambda_n \geq \frac{(1-\alpha)2^{h+1}((\lambda^*_n)^2+\sigma^2)}{(2^{h+1}+1)\sigma^2} \sim \frac{2(1-\alpha)q\log n}{1+\gamma}. \tag{4.14}$$

Assuming (4.10), using (4.8), and for $n \geq n_0(K(\gamma)+1)$, (4.12) becomes

$$\begin{aligned}\Lambda_n &\geq \frac{1}{\sigma^2}\sum_{j\leq h,k} P(|z_{j,k}| \geq K(\gamma)+1)\\ &\geq \frac{\alpha 2^h}{2\sigma^2}\left(1-\Phi\left(\frac{K(\gamma)+1)}{\sigma}\right)\right) \sim C_4 2^h \geq C_4\frac{n^q}{2},\end{aligned} \tag{4.15}$$

where the sum is over the indices such that $\lambda^*_{j,k,n}$ satisfy (4.10), where $C_4 = C_4(\alpha,\sigma,K(\gamma))$ and since $n^q \leq 2^{h+1} \leq 2n^q$. Finally, assuming (4.11), (4.12) becomes

$$\begin{aligned}\Lambda_n &\geq \sum_{j\leq h,k}\frac{((\lambda^*_{j,k,n})^2+\sigma^2)}{(2^{h+1}+1)\sigma^2}\\ &\geq \frac{\alpha 2^h}{(2^{h+1}+1)\sigma^2}\left(\frac{\varepsilon(\gamma)\sigma^2 2^{(m-h)/2}}{C_3 M}+\sigma^2\right)^2 \sim C_5 2^{(m-h)} \geq C_5 n^{1-q},\end{aligned} \tag{4.16}$$

where the sum is over the $\lambda^*_{j,k,n}$ satisfying (4.11) and for some appropriate constant C_5. Thus (4.14)–(4.16) show that the right-hand side of (4.4) grows as least as fast as $(1-\alpha)2q\log n/(1+\gamma)$. Since α, q and γ are arbitrary, this gives

$$\liminf_{n\to+\infty}\frac{\Lambda_n}{2\log n} \geq 1. \qquad \square$$

5. Smooth compactly supported densities. In this section we show that soft thresholding is optimal for a class of C^2-noise with compactly supported positive density. Under these conditions on the noise, no estimator can give a better rate than soft thresholding. However, in the "function space setting" where signals are assumed to belong to balls in function spaces, and without the "smoothness" assumption, other types of estimators can outperform soft thresholding. Such an improvement is presented, via a nonlinear filtering procedure, and for uniform noise, in Section 4 of our companion paper [3]. Moreover, estimators which exploit the special structure of the densities involved, for example a moving median estimator (see Section 3 of [3]), can also outperform soft thresholding for other types of noise, e.g., for noise with inverse polynomial tails.

THEOREM 5.1. *Let the observations $X_i = s_i + e_i$, $i = 1,\dots,n = 2^m$, be given, where the s_i are parameters of interest and the e_i are zero mean i.i.d. random variables with variance σ^2 and law μ. Let μ be absolutely continuous with compactly supported (on $[a,b]$) density f. Let f be twice differentiable on $[a,b]$ and such that $f>0$ on (a,b). Further, let $f(a)=f(b)=f'(a)=f'(b)=0$, let $f''(a)\neq 0$, $f''(b)\neq 0$ and let $\sup_{a<x<b}|\sqrt{f}''(x)|<+\infty$. Apply a periodic wavelet transform W_n to these observations, taking for every n the same compactly*

supported generating wavelet. Let $Y = W_n(X)$ *and let* $\theta = W_n(s)$. *Then*

$$\liminf_{n\to\infty} \inf_{\widehat{\theta}} \sup_{\theta\in\mathbb{R}^n} \frac{E\|\widehat{\theta}-\theta\|^2}{B_n(\theta,\sigma^2)} \frac{1}{\log n} > 0,$$

where the infimum is taken over all the estimators of θ.

The computations of the minimax bounds in the previous sections were based on the a priori measure $F_{\varepsilon,a}$ and its Bayes risk, where ε and a were carefully chosen. In that context we computed the Bayes estimator using likelihood ratios. Here we wish to apply a similar method to coarse level wavelet coefficients. Again the Bayes estimator relies on likelihood ratios, and below we compute an asymptotic expansion for these likelihood ratios. To do so, we will use elements of Le Cam–Hájek LAN theory as exposed in [37], for example.

PROOF OF THEOREM 5.1. Let m be the Lebesgue measure, let $P_0 = fm$ and let $P_h = P_0 * \delta_h$, that is, $P_h(dx) = f(x-h)m(dx)$ and let also $g := -2\sqrt{f}'/\sqrt{f} = -f'/f$ on (a,b), $g = 0$ elsewhere. Finally, let $t_{n,i}$, $i = 1,\dots,n$, be a triangular array of numbers with $t_n := \sup_{i=1,\dots,n} |t_{n,i}|$ and $t_n/n^c \to 0$, for any $c > 0$. Our goal is to prove that

$$\log\left(\prod_{i=1}^n \frac{dP_{t_{n,i}/\sqrt{n}}}{dP_0}(x_i)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n t_{n,i}\,g(x_i) - \frac{E_{P_0}|g|^2}{2n}\sum_{i=1}^n t_{n,i}^2 + o_{P_0^n}(1), \tag{5.1}$$

where $o_{P_0^n}(1)$ denotes a sequence of random variables converging in probability to 0, and where this stochastic convergence only depends on t_n and n. For simplicity and ease of notation we will assume that $t_{n,i} \geq 0$, as this spares us some simple distinction of cases. For $2 \leq p < 3$, let $E_{P_0}|g|^p = \int_a^b |f'(x)|^p/f(x)^{p-1}\,dx$. The restrictions on f imply that for $x \in (a,b)$ close to a, $f(x) = c(x-a)^2 + o((x-a)^2)$ and $f'(x) = 2c(x-a) + o((x-a))$, for some constant $c \neq 0$. Thus, near a, $|f'(x)|^p/f(x)^{p-1} = O((x-a)^{-(p-2)})$. A similar statement holds for f close to b, and together these imply that $E_{P_0}|g|^p < \infty$. Let us now turn to the likelihood quotient, and define

$$h_{n,i} := 2\left(\sqrt{\frac{dP_{t_{n,i}/\sqrt{n}}}{dP_0}} - 1\right).$$

For $2 \leq p < 3$, then

$$\int_{\mathbb{R}} |h_{n,i}|^p\,dP_0 = 2^p \int_a^b \left|\sqrt{f(x-t_{n,i}/\sqrt{n})} - \sqrt{f(x)}\right|^p f(x)^{1-p/2}\,dx < +\infty. \tag{5.2}$$

Indeed, the conditions on f imply that $\sqrt{f} \in C^1([a,b])$, hence

$$\left|\sqrt{f(x-t_{n,i}/\sqrt{n})} - \sqrt{f(x)}\right| = O(t_n/\sqrt{n}), \tag{5.3}$$

uniformly in x. Moreover, since near a (resp. near b), $f(x)^{1-p/2} = O((x-a)^{2-p})$ [resp., $f(x)^{1-p/2} = O((b-x)^{2-p})$], it follows that $\int_a^b f(x)^{1-p/2}\,dx < \infty$. Thus, $E|h_{n,i}|^p < \infty,\ 2 \le p < 3$, and in fact

$$E_{P_0}|h_{n,i}|^p = O\left(\left(\frac{t_n}{\sqrt{n}}\right)^p\right). \tag{5.4}$$

Next, we show that

$$E_{P_0}\left|h_{n,i} - \frac{t_{n,i}}{\sqrt{n}}g\right|^2 = \int_{-\infty}^{\infty}\left(h_{n,i}(x) - \frac{t_{n,i}}{\sqrt{n}}g(x)\right)^2 f(x)\,dx = O\left(\frac{t_n^3}{n^{3/2}}\right). \tag{5.5}$$

The middle term in (5.5) is equal to

$$\begin{aligned}
&4\int_a^b \left(\frac{\sqrt{f(x - t_{n,i}/\sqrt{n})}}{\sqrt{f(x)}} - 1 - \frac{t_{n,i}}{2\sqrt{n}}g(x)\right)^2 f(x)\,dx \\
&\quad = 4\int_a^b \left(\sqrt{f(x - t_{n,i}/\sqrt{n})} - \sqrt{f(x)} + \frac{t_{n,i}}{\sqrt{n}}\sqrt{f(x)}'\right)^2 dx \\
&\quad = 4\int_a^{a+t_{n,i}/\sqrt{n}} \left(\sqrt{f(x - t_{n,i}/\sqrt{n})} - \sqrt{f(x)} + \frac{t_{n,i}}{\sqrt{n}}\sqrt{f(x)}'\right)^2 dx \\
&\qquad + 4\int_{a+t_{n,i}/\sqrt{n}}^b \left(\sqrt{f(x - t_{n,i}/\sqrt{n})} - \sqrt{f(x)} + \frac{t_{n,i}}{\sqrt{n}}\sqrt{f(x)}'\right)^2 dx \\
&\quad \le 8\int_a^{a+t_{n,i}/\sqrt{n}} \left(\left(\sqrt{f(x - t_{n,i}/\sqrt{n})} - \sqrt{f(x)}\right)^2 + \frac{t_{n,i}^2}{n}\sqrt{f(x)}'^2\right) dx \\
&\qquad + 4\int_{a+t_{n,i}/\sqrt{n}}^b \left(c_1\frac{t_{n,i}}{\sqrt{n}}\right)^4 dx \qquad \left(\text{since } \sup_{a<x<b}|\sqrt{f}''(x)| < +\infty\right) \\
&\quad \le c_2\left(\left(\frac{t_n}{\sqrt{n}}\right)^3 + \left(\frac{t_n}{\sqrt{n}}\right)^3 + \frac{t_n^4}{n^2}\right),
\end{aligned}$$

for some constants $c_1 > 0, c_2 > 0$, by (5.3) and since $\sqrt{f} \in C^1([a,b])$. This proves (5.5).

As in [37], page 379, a simple Taylor expansion gives

$$\begin{aligned}
&\log\left(\prod_{i=1}^n \frac{dP_{t_{n,i}/\sqrt{n}}}{dP_0}(x_i)\right) \\
&\quad = \sum_{i=1}^n 2\log\left(\frac{1}{2}h_{n,i}(x_i) + 1\right) \qquad (5.6) \\
&\quad = \sum_{i=1}^n h_{n,i}(x_i) - \frac{1}{4}\sum_{i=1}^n h_{n,i}(x_i)^2 + \frac{1}{4}\sum_{i=1}^n \big(1 - r(h_{n,i}(x_i))\big)h_{n,i}(x_i)^2,
\end{aligned}$$

where $r(0) = 1$ and where $|r(x_1) - r(x_2)| \le C|x_1 - x_2|$, for $|x_1| < 1$, $|x_2| < 1$ and some constant $C > 0$. We now analyze (5.6), and first show that the last summand there converges to zero in probability, the rate depending only on t_n and n. First,

$$
\begin{aligned}
&P_0^n\left(\sum_{i=1}^n \bigl(1 - r(h_{n,i}(x_i))\bigr)h_{n,i}(x_i)^2 \ge \varepsilon\right)\\
&\quad\le P_0^n\left(\sup_i |1 - r\bigl(h_{n,i}(x_i)\bigr)| \ge \frac{1}{t_n^2 \log n}\right)\\
&\qquad + P_0^n\left(\frac{1}{t_n^2 \log n}\sum_{i=1}^n h_{n,i}(x_i)^2 \ge \varepsilon\right).
\end{aligned}
\tag{5.7}
$$

Next, by (5.4), $E_{P_0}|h_{n,i}|^2 = O(t_n^2/n)$, hence $\frac{1}{t_n^2 \log n}\sum_{i=1}^n h_{n,i}(x_i)^2$ converges to zero in L^1, and the rightmost term in (5.7) converges to zero in probability, the rate depending only on n. Further for $2 < p < 3$,

$$
\begin{aligned}
&P_0^n\left(t_n^2 \log n \sup_i |1 - r(h_{n,i}(x_i))| \ge 1\right)\\
&\quad\le \sum_{i=1}^n P_0\left(|1 - r(h_{n,i}(x_i))| \ge \frac{1}{t_n^2 \log n}\right)\\
&\quad\le \sum_{i=1}^n P_0\left(|h_{n,i}(x_i)| \ge \frac{1}{C t_n^2 \log n}\right)\\
&\quad\le \sum_{i=1}^n E_{P_0}|h_{n,i}|^p C^p t_n^{2p} (\log n)^p\\
&\quad= O\left(\frac{t_n^{3p}(\log n)^p}{n^{p/2-1}}\right),
\end{aligned}
$$

where the second inequality follows from properties of r and the last using (5.4). Since $p > 2$, our choice of t_n implies the uniform stochastic convergence of (5.7), and thus

$$
\log\left(\prod_{i=1}^n \frac{dP_{t_{n,i}/\sqrt{n}}}{dP_0}(x_i)\right) = \sum_{i=1}^n h_{n,i}(x_i) - \frac{1}{4}\sum_{i=1}^n h_{n,i}(x_i)^2 + o_{P_0^n}(1).
\tag{5.8}
$$

Next, we show that $\sum_{i=1}^n h_{n,i}^2$ and $\sum_{i=1}^n E_{P_0} h_{n,i}^2$ are stochastically equivalent with respect to P_0^n, that is, that $\sum_{i=1}^n X_{n,i} \to 0$ in probability, where $X_{n,i} = h_{n,i}^2 - E_{P_0} h_{n,i}^2$. For $2 < p < 3$, (5.4) gives

$$
E_{P_0}|X_{n,i}|^{p/2} \le E_{P_0}|h_{n,i}|^p + (E_{P_0} h_{n,i}^2)^{p/2} = O\left(\frac{t_n^p}{n^{p/2}}\right),
$$

hence if $\widetilde{X}_{n,i} := X_{n,i}\mathbf{1}_{\{|X_{n,i}|\le 1\}}$, then $\widetilde{X}_{n,i}^2 \le |X_{n,i}|^{p/2}$ and

$$\sum_{i=1}^{n} E_{P_0}\widetilde{X}_{n,i}^2 \le \sum_{i=1}^{n} E_{P_0}|X_{n,i}|^{p/2} = O\Big(\frac{t_n^p}{n^{p/2-1}}\Big). \tag{5.9}$$

Since $E_{P_0}X_{n,i} = 0$,

$$\begin{aligned}\sum_{i=1}^{n} \big|E_{P_0}X_{n,i}\mathbf{1}_{\{|X_{n,i}|\le 1\}}\big| &\le \sum_{i=1}^{n} E_{P_0}|X_{n,i}|\mathbf{1}_{\{|X_{n,i}|>1\}} \\ &\le \sum_{i=1}^{n}\Big(1P_0(|X_{n,i}|\ge 1) + \int_1^\infty \frac{E_{P_0}|X_{n,i}|^{p/2}}{x^{p/2}}\,dx\Big) \qquad (5.10)\\ &= O\Big(\frac{t_n^p}{n^{p/2-1}}\Big),\end{aligned}$$

by (5.9). Now, for $\varepsilon > 0$,

$$P_0^n\Big(\Big|\sum_{i=1}^{n} X_{n,i}\Big| \ge \varepsilon\Big) \le \sum_{i=1}^{n} P_0(|X_{n,i}| > 1) + P_0^n\Big(\Big|\sum_{i=1}^{n}\widetilde{X}_{n,i}\Big| \ge \varepsilon\Big)$$

and

$$\sum_{i=1}^{n} P_0(|X_{n,i}|\ge 1) \le \sum_{i=1}^{n} E_{P_0}|X_{n,i}|^{p/2} = O\Big(\frac{t_n^p}{n^{p/2-1}}\Big),$$

while

$$\begin{aligned}P_0^n\Big(\Big|\sum_{i=1}^{n}\widetilde{X}_{n,i}\Big| \ge \varepsilon\Big) &\le \frac{E_{P_0^n}(\sum_{i=1}^{n}\widetilde{X}_{n,i})^2}{\varepsilon^2} \\ &\le \frac{\sum_{i=1}^{n} E_{P_0}\widetilde{X}_{n,i}^2 + (\sum_{i=1}^{n}|E_{P_0}\widetilde{X}_{n,i}|)^2}{\varepsilon^2} \\ &= O\Big(\frac{t_n^p}{n^{p/2-1}}\Big),\end{aligned}$$

by (5.9) and (5.10). It thus follows that $\sum_{n,i} X_{n,i}$ converges to zero uniformly, that is, the rate depends only on t_n and n. Combining the above with (5.8) gives

$$\begin{aligned}&\log\Big(\prod_{i=1}^{n}\frac{dP_{t_{n,i}/\sqrt{n}}}{dP_0}\Big) \\ &= \sum_{i=1}^{n} h_{n,i} - \frac{1}{4}\sum_{i=1}^{n} h_{n,i}^2 + o_{P_0^n}(1) \qquad (5.11)\\ &= \sum_{i=1}^{n}(h_{n,i} - E_{P_0}h_{n,i}) - \frac{1}{4}\sum_{i=1}^{n} E_{P_0}h_{n,i}^2 + \sum_{i=1}^{n} E_{P_0}h_{n,i} + o_{P_0^n}(1).\end{aligned}$$

Still proceeding as in [37], page 381 top,

$$E_{P_0} h_{n,i} = -\tfrac{1}{4} E_{P_0} h_{n,i}^2 - P_{t_{n,i}/\sqrt{n}}(N_{n,i}),$$

where $N_{n,i} = \{dP_0/dP_{t_{n,i}/\sqrt{n}} = 0\}$. Hence (5.11) becomes

$$\begin{aligned} &\log\left(\prod_{i=1}^n \frac{dP_{t_{n,i}/\sqrt{n}}}{dP_0}(x_i)\right) \\ &\quad = \sum_{i=1}^n (h_{n,i} - E_{P_0} h_{n,i}) - \frac{1}{2}\sum_{i=1}^n E_{P_0} h_{n,i}^2 - \sum_{i=1}^n P_{t_{n,i}/\sqrt{n}}(N_{n,i}) + o_{P_0^n}(1). \end{aligned} \tag{5.12}$$

Next,

$$\begin{aligned} &\sum_{i=1}^n \left| E_{P_0}\left(\frac{t_{n,i}}{\sqrt{n}}\, g\right)^2 - E_{P_0}(h_{n,i})^2 \right| \\ &\quad = \sum_{i=1}^n \left| E_{P_0}\left(\frac{t_{n,i}}{\sqrt{n}}\, g - h_{n,i}\right)\left(\frac{t_{n,i}}{\sqrt{n}}\, g + h_{n,i}\right) \right| \\ &\quad \le \sum_{i=1}^n \sqrt{E_{P_0}\left(\frac{t_{n,i}}{\sqrt{n}}\, g - h_{n,i}\right)^2 E_{P_0}\left(\frac{t_{n,i}}{\sqrt{n}}\, g + h_{n,i}\right)^2} \\ &\quad = O\left(n\sqrt{\left(\frac{t_n^3}{n^{3/2}}\right)\left(\frac{t_n^2}{n}\right)}\right) = O\left(\frac{t_n^{5/2}}{\sqrt[4]{n}}\right), \end{aligned} \tag{5.13}$$

by (5.4) and (5.5). Moreover, since $E_{P_0} g = 0$,

$$\begin{aligned} &E_{P_0^n}\left(\sum_{i=1}^n (h_{n,i} - E_{P_0} h_{n,i}) - \frac{t_{n,i}}{\sqrt{n}}\, g\right)^2 \\ &\quad = \sum_{i=1}^n E_{P_0^n}\left(h_{n,i} - \frac{t_{n,i}}{\sqrt{n}}\, g - E_{P_0} h_{n,i}\right)^2 \\ &\quad \le \sum_{i=1}^n E_{P_0^n}\left(h_{n,i} - \frac{t_{n,i}}{\sqrt{n}}\, g\right)^2 = O\left(\frac{t_n^3}{\sqrt{n}}\right), \end{aligned} \tag{5.14}$$

by (5.5). Hence with (5.13) and (5.14), (5.12) becomes

$$\begin{aligned} &\log\left(\prod_{i=1}^n \frac{dP_{t_{n,i}/\sqrt{n}}}{dP_0}(x_i)\right) \\ &\quad = \sum_{i=1}^n \frac{t_{n,i}}{\sqrt{n}} g(x_i) - \frac{1}{2}\sum_{i=1}^n \left(\frac{t_{n,i}}{\sqrt{n}}\right)^2 E_{P_0} g^2 - \sum_{i=1}^n P_{t_{n,i}/\sqrt{n}}(N_{n,i}) + o_{P_0^n}(1), \end{aligned}$$

where again the stochastic convergence depends only on t_n and n. Finally, since

$$P_{t_{n,i}/\sqrt{n}}(N_{n,i}) = \int_b^{b+t_{n,i}/\sqrt{n}} f\left(x - \frac{t_{n,i}}{\sqrt{n}}\right) dx$$
$$= \int_0^{t_{n,i}/\sqrt{n}} f(b-x)\,dx = O\left(\frac{t_n^3}{n^{3/2}}\right),$$

by the conditions imposed on f, it follows that

$$\log \prod_{i=1}^n \frac{dP_{t_{n,i}/\sqrt{n}}}{dP_0}(x_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^n t_{n,i}\, g(x_i) - \frac{E_{P_0} g^2}{2n} \sum_{i=1}^n t_{n,i}^2 + o_{P_0^n}(1),$$

where once more the stochastic convergence of $o_{P_0^n}(1)$ to 0 depends only on n and t_n. As announced in (5.11), we obtained an asymptotic expansion for the likelihood ratios which we are going to use to compute Bayes risks.

Let us now come to the second part of the proof. Again, we only consider the wavelet coefficients at the level $l = (\log_2 n)/2$, the coefficients in the middle of the wavelet coefficients pyramid. The Bayes measure for each $\theta_{j,k}$, $j \neq l$, is δ_0, while for $j = l$ it is F_{ε_n, a_n}, with $\varepsilon_n \to 0$ and $a_n \to \infty$; the exact sequences will be specified later. The overall Bayes measure is the product measure of the individual Bayes measures. First let us compute the Bayes risk for a single coefficient $\theta_{l,k}$. Let $z = W(e)$, and let $\widetilde{\theta}_{j,k}$ be random variables which are distributed according to the Bayes measure, and which are independent of the initial noise (e_i). The Bayes estimator for the coefficient with index (l,h) is

$$E\big(\widetilde{\theta}_{l,h} | z_{j,k} + \widetilde{\theta}_{j,k},\ j = 0, \ldots, m-1;\ k = 1, \ldots, 2^j\big),$$

and the Bayes risk of estimation is (the expectation is for the noise and the Bayes measure simultaneously)

$$E\big(E\big(\widetilde{\theta}_{l,h} | z_{j,k} + \widetilde{\theta}_{j,k},\ j = 0, \ldots, m-1;\ k = 1, \ldots, 2^j\big) - \widetilde{\theta}_{l,h}\big)^2$$
$$\geq E\big(E\big(\widetilde{\theta}_{l,h} | z_{j,k} + \widetilde{\theta}_{j,k},\ j = 0, \ldots, m-1;\ k = 1, \ldots, 2^j;$$
$$\widetilde{\theta}_{j,k}, (j,k) \neq (l,h),\ j = 0, \ldots, m-1; k = 1, \ldots, 2^j\big) - \widetilde{\theta}_{l,h}\big)^2$$
$$= E\big(E\big(\widetilde{\theta}_{l,h} | z_{l,h} + \widetilde{\theta}_{l,h},\ z_{j,k},\ (j,k) \neq (l,h),$$
$$j = 0, \ldots, m-1;\ k = 1, \ldots, 2^j\big) - \widetilde{\theta}_{l,h}\big)^2$$
$$\geq E\big(E(\widetilde{\theta}_{l,h} | e_i + \widetilde{\theta}_{l,h} c_{l,h,i}, i = 1, \ldots, n) - \widetilde{\theta}_{l,h}\big)^2,$$

where the last inequality holds since $\sum_i (e_i + \widetilde{\theta}_{l,h} c_{l,h,i}) c_{l,h,i} = z_{l,h} + \widetilde{\theta}_{l,h}$ and for $(j,k) \neq (l,h)$, $\sum_i (e_i + \widetilde{\theta}_{l,h} c_{l,h,i}) c_{j,k,i} = z_{j,k}$, where $(c_{j,k,i})$ are the coefficients of the wavelet transform (recall that the wavelet transform is orthonormal). Now, simple computations yield that

$$E(\widetilde{\theta}_{l,h} | e_\cdot + \widetilde{\theta}_{l,h} c_{l,h,\cdot}) = a_n \frac{\varepsilon_n\, dP_{a_n}}{\varepsilon_n\, dP_{a_n} + (1-\varepsilon_n)\, dP_0},$$

where P_0 is the law of $e_\cdot$ and dP_{a_n} the law of $e_\cdot + a_n(c_{l,h,\cdot})$. With the background of the proofs of Theorems 3.1 and 3.4 it is easy to see that when choosing $\varepsilon_n = \log(\sqrt{n})/\sqrt{n}$, the Bayes risk for this coefficient is larger than

$$(1-\varepsilon_n)^2 \varepsilon_n a_n^2 \int_{\mathbb{R}^n} \left(\frac{dP_0}{(1-\varepsilon_n)\,dP_0 + \varepsilon_n\,dP_{a_n}} \right)^2 dP_{a_n}.$$

Let p_1, C_1 be constants between 0 and 1. If the above integrand is larger than C_1 with probability p_1 then the Bayes risk is larger than $(1-\varepsilon_n)^2 \varepsilon_n a_n^2 C_1 p_1$. Now, since

$$\begin{aligned} &P_{a_n}\left(\left(\frac{dP_0}{(1-\varepsilon_n)\,dP_0 + \varepsilon_n\,dP_{a_n}} \right)^2 \ge C_1 \right) \\ &\quad = P_0\left(\left(\frac{dP_{-a_n}}{(1-\varepsilon_n)\,dP_{-a_n} + \varepsilon_n\,dP_0} \right)^2 \ge C_1 \right) \\ &\quad = P_0\left(\left(1 - \varepsilon_n + \frac{\varepsilon_n\,dP_0}{dP_{-a_n}} \right)^2 \le 1/C_1 \right) \\ &\quad \ge P_0\left(\frac{dP_{-a_n}}{dP_0} \ge C_2 \varepsilon_n \right), \end{aligned} \tag{5.15}$$

where $C_2 = 1/(1/\sqrt{C_1} - 1)$ and since clearly the asymptotic properties of dP_{-a_n}/dP_0 and dP_{a_n}/dP_0 are the same. Let us investigate when $P_0(dP_{a_n}/dP_0 > C_2\varepsilon_n) > p_1$. We have $Y_{l,h} = \sum_{i=1}^n c_{l,h,i} X_i$. Since $l = (\log_2 n)/2$, and thanks to the dilation equation, only about $r = O(\sqrt{n})$ of the $c_{l,h,i}^n$ are nonzero (see [8], Chapters 6–8]). Thus, if c_i is short for the nonzero coefficients $c_{l,h,i}$, and re-indexing X_i as Z_i, we get $Y_{l,h} = \sum_{i=1}^r c_i Z_i$, with $\sup_i c_i^2 \le C_3/r$, where C_3 is some global constant, as we already know from the proof of Theorem 4.1. Next, changing the mean of the $Y_{l,h}$ by a is equivalent to changing the mean of each Z_i by ac_i. [It follows from the orthonormality of the wavelet transform that the inverse wavelet transform of $(w_{j,k})$ where $w_{l,h} = a$ and $w_{j,k} = 0$ for $(j,k) \ne (l,h)$ is $(a_{l,h,i})_i$.] It thus follows from (5.1) (with $t_{n,i} = a_n c_i \sqrt{r}$) that

$$\frac{dP_{a_n}}{dP_0} = \exp\left(a_n U_n - \frac{a_n^2}{2}\gamma^2 + o_{P_0}(1) \right), \tag{5.16}$$

where $U_n = \frac{1}{\sqrt{r}} \sum_{i=1}^r (c_i\sqrt{r}) g$ and $\gamma^2 = E_{P_0} g^2$. Note that $\sum_{i=1}^r (\sqrt{r} c_i)^2 = r$. By the central limit theorem, U_n converges in distribution to a $N(0, \gamma^2)$ random variable, that is, $U_n = V_n + R_n$, where V_n is a $N(0,\gamma^2)$ random variable and R_n converges to 0 in probability. Since $E|g(X_i)|^p, 2 < p < 3$, the uniform convergence in the central limit theorem ensures that this convergence only depends on $t_n := \sup_i |t_{n,i}|$ and n (e.g., see Theorem 5.8 in [34]). Thus from (5.15)

and (5.16), we wish to investigate when

$$P_0\Big(V_n + o_{P_0}(1) \geq \frac{\log(\varepsilon_n C_2)}{a_n} + \frac{a_n\gamma^2}{2}\Big) > p_1.$$

Let $\varepsilon_n := (\log\sqrt{n})/\sqrt{n}$, and let a_n be the solution (which exists as shown below) of

$$\frac{\log(\varepsilon_n C_2)}{a_n} + \frac{a_n\gamma^2}{2} + \gamma = \gamma\,\Phi^{-1}\Big(\frac{1-p_1}{2}\Big), \tag{5.17}$$

where Φ is the standard normal distribution function. Now for some n_0 and all $n \geq n_0$,

$$\begin{aligned} &P_0\Big(V_n + o_{P_0}(1) \geq \frac{\log(\varepsilon_n C_2)}{a_n} + \frac{a_n\gamma^2}{2}\Big) \\ &\quad\geq P_0\Big(V_n \geq \frac{\log(\varepsilon_n C_2)}{a_n} + \frac{a_n\gamma^2}{2} + \gamma\Big) - \frac{1-p_1}{2} \\ &\quad= 1 - \Phi\Big(\frac{\log(\varepsilon_n C_2)}{\gamma a_n} + \frac{a_n\gamma}{2} + 1\Big) - \frac{1-p_1}{2} = p_1, \end{aligned}$$

by (5.17). Next, let us provide a closed form expression for a_n and find its asymptotics. Simple computations yield that the solution of (5.17) is

$$\begin{aligned} a_n &= \sqrt{\Big(\frac{1-\Phi^{-1}((1-p_1)/2)}{\gamma}\Big)^2 - \frac{2\log(\varepsilon_n C_2)}{\gamma^2}} - \frac{1-\Phi^{-1}((1-p_1)/2)}{\gamma} \\ &= \sqrt{\frac{\log n}{\gamma^2} - 2\frac{\log C_2 + \log\log\sqrt{n}}{\gamma^2} + \Big(\frac{1-\Phi^{-1}((1-p_1)/2)}{\gamma}\Big)^2} \\ &\quad - \frac{1-\Phi^{-1}((1-p_1)/2)}{\gamma} \\ &\sim \frac{\sqrt{\log n}}{\gamma}. \end{aligned}$$

These choices of ε_n and a_n provide lower bounds on the Bayes risks and applying the machinery of the proofs of Theorems 3.1 and 3.4 gives

$$\liminf_{n\to\infty} \inf_{\widehat{\theta}} \sup_{\theta} \frac{E\|\theta - \widehat{\theta}\|^2}{\sigma^2 + \sum_{i=1}^n \min(\theta_i^2, \sigma^2)} \frac{\gamma^2}{\log n} > 0. \qquad \square$$

6. Concluding remarks. Variations of many of our results, such as Theorems 2.1 and 2.7 and Proposition 2.5 also hold for related types of estimators,

for example, hard thresholding or the estimator $T_\lambda^M(x) := x\mathbf{1}_{\{|x|\geq\lambda\}} + 2(|x| - \lambda/2)_+ \operatorname{sgn}(x)\mathbf{1}_{\{|x|<\lambda\}}$, which belongs to the semisoft class of [6]. In each case, the size of the threshold parameter and the performance of the estimator are governed by the tail behavior of the noise distribution. The asymptotic performance is the same as for soft thresholding, as long as the density satisfies the conditions of Theorem 2.7, that is, it decays like $\exp(-h(x))$, where h grows at least as fast as x^ε, $\varepsilon > 0$.

If applied levelwise, the ideal estimator method is no longer minimax. The thresholds are a little bit too small, the sum of the risks at 0 of the coefficients is of size $\sim C\log^2 n$, for some constant C. In practice the coarser levels are not thresholded, and this does not change the asymptotic performance of the method since only a smaller and smaller fraction of the wavelet coefficients is not thresholded. Another method, where the threshold for the kth coefficient is always the same (no matter what the length of the input vector is) is briefly introduced now in the normal case. The result below can be transfered to other exponentially decaying densities as in Theorem 2.7; the main point is to choose $\widetilde{\lambda}_i$ such that $\sum_{i=1}^n p(\widetilde{\lambda}_i, 0) \approx \Lambda_n\sigma^2$.

THEOREM 6.1. *Let $Y_i = \theta_i + z_i$, $i = 1, \ldots, n$, where the z_i are i.i.d. normal random variables with mean zero and variance σ^2, and let Λ_n and $p(\cdot,\cdot)$ have their usual meaning (as in Theorem 2.1). Let $\widetilde{\lambda}_i$ be such that $p(\widetilde{\lambda}_i, 0) = 2\sigma^2/i$, and let*

$$\widetilde{\Lambda}_n := \sup_{\theta\in\mathbb{R}^n} \frac{E\sum_{i=1}^n |T_{\widetilde{\lambda}_i}^S(Y_i) - \theta_i|^2}{B_n(\theta, \sigma^2)}.$$

Then $\lim_{n\to\infty} \Lambda_n/\widetilde{\Lambda}_n = 1$.

PROOF. First, let us provide asymptotics for $\widetilde{\lambda}_n$. To do so, recall (e.g., see [35], page 850) that as $\lambda \to +\infty$,

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_\lambda^\infty e^{-x^2/(2\sigma^2)}\,dx \sim \frac{1}{\sqrt{2\pi}} \frac{\sigma}{\lambda} \exp\left(-\frac{\lambda^2}{2\sigma^2}\right). \tag{6.1}$$

Let

$$s(\lambda) := p(\lambda, 0) = \frac{2}{\sqrt{2\pi\sigma^2}} \int_\lambda^\infty (x-\lambda)^2 e^{-x^2/(2\sigma^2)}\,dx;$$

clearly s is continuous, increasing and let s^{-1} denote its inverse. From (6.1), it is easily seen that $\lim_{x\to 0} s^{-1}(x)/\sqrt{2\sigma^2\log(1/x)} = 1$. Our choice of thresholds $[p(\widetilde{\lambda}_i, 0) = 2\sigma^2/i]$ implies that

$$\lim_{n\to\infty} \frac{\widetilde{\lambda}_n}{\sqrt{2\sigma^2\log n}} = 1. \tag{6.2}$$

Taking in the defining property of $\widetilde{\Lambda}_n$, $\theta_i = 0$ for all i, we get $\widetilde{\Lambda}_n \geq (\sum_{i=1}^n p(\widetilde{\lambda}_i, 0))/\sigma^2 = \sum_{i=1}^n 2/i$. Since $\Lambda_n \sim 2\log n$, it follows that $\limsup_{n\to+\infty} \Lambda_n/\widetilde{\Lambda}_n \leq 1$. In particular, $\widetilde{\Lambda}_n \geq 1$. Let $\theta \in \mathbb{R}^n$, if $|\theta_j| \geq \sigma$. Then since $p(\widetilde{\lambda}_j, \theta_j) \leq p(\widetilde{\lambda}_j, \infty)$ (see the proof of Theorem 2.1),

$$\frac{\sum_{i=1}^n p(\widetilde{\lambda}_i, \theta_i)}{B_n(\theta, \sigma^2)} \leq \frac{p(\widetilde{\lambda}_j, \infty) + \sum_{i=1, i\neq j}^n p(\widetilde{\lambda}_i, \theta_i)}{B_n(\theta, \sigma^2)}. \tag{6.3}$$

If $|\theta_j| \leq \sigma$ and if, moreover, $\sum_{i=1}^n p(\widetilde{\lambda}_i, \theta_i)/B_n(\theta, \sigma^2) \geq 1$, then by (2.8),

$$\begin{aligned} \frac{\sum_{i=1}^n p(\widetilde{\lambda}_i, \theta_i)}{B_n(\theta, \sigma^2)} &\leq \frac{p(\widetilde{\lambda}_j, 0) + \theta_j^2 + \sum_{i=1, i\neq j}^n p(\widetilde{\lambda}_i, \theta_i)}{\sigma^2 + \theta_j^2 + \sum_{i=1, i\neq j}^n \min(\theta_i^2, \sigma^2)} \\ &\leq \frac{p(\widetilde{\lambda}_j, 0) + \sum_{i=1, i\neq j}^n p(\widetilde{\lambda}_i, \theta_i)}{\sigma^2 + \sum_{i=1, i\neq j}^n \min(\theta_i^2, \sigma^2)}. \end{aligned} \tag{6.4}$$

Since $\sup_{\theta\in\mathbb{R}^n} \frac{\sum_{i=1}^n p(\widetilde{\lambda}_i, \theta_i)}{B_n(\theta, \sigma^2)} \geq 1$, repeated use ($n$ times) of (6.3) and (6.4) leads to

$$\begin{aligned} \sup_{\theta\in\mathbb{R}^n} \frac{\sum_{i=1}^n p(\widetilde{\lambda}_i, \theta_i)}{B_n(\theta, \sigma^2)} &\leq \sup_{\theta\in\{0,\infty\}^n} \frac{\sum_{i=1}^n p(\widetilde{\lambda}_i, \theta_i)}{B_n(\theta, \sigma^2)} \\ &= \sup_{J\subset\{1,\dots,n\}} \frac{\sum_{i\in J} p(\widetilde{\lambda}_i, 0) + \sum_{i\in J^c} (\widetilde{\lambda}_i^2 + \sigma^2)}{\sigma^2(1 + |J^c|)} \\ &\leq \sup_{J\subset\{1,\dots,n\}} \frac{(2 + 2\log n)\sigma^2 + |J^c|(\widetilde{\lambda}_n^2 + \sigma^2)}{\sigma^2(1 + |J^c|)} \sim \Lambda_n, \end{aligned}$$

since $\sum_{i=1}^n 1/i \leq 1 + \log n$, since $\widetilde{\lambda}_i$ is increasing with i and since from (6.2), $(\widetilde{\lambda}_n^2 + \sigma^2)/\sigma^2 \sim \Lambda_n \sim 2\log n$. Hence, $\liminf_{n\to+\infty} \Lambda_n/\widetilde{\Lambda}_n \geq 1$, and the proof is complete. □

It is known that the ideal estimator is not always the optimal one. Indeed let X be a zero mean random variable with variance 1 and let $x \to \alpha x$, $\alpha \in (0, 1)$ be a linear shrinker. Then $E(\alpha(X+\theta) - \theta)^2 = \alpha^2 + (1-\alpha)^2\theta^2$. If $|\theta| = 1$ and $\alpha = 1/2$ then $((1-\alpha)^2\theta^2 + \alpha^2)/(\min(\theta^2, 1)) = 1/2$. This pathology is because we are estimating a single coefficient whose square is the variance of the noise. However, in general, linear shrinkers applied to wavelet coefficients do not always perform that well. Indeed, a result of Donoho and Johnstone ([14], Theorem 5) asserts that for $X_1, \dots, X_n$, i.i.d. centered normal random variables with variance σ^2,

$$\inf_\alpha E\|\alpha(X+\theta) - \theta\|^2 \geq E\|T_{\mathrm{JS}}(X+\theta) - \theta\|^2 - 2\sigma^2,$$

for all $\theta \in \mathbb{R}^n$, where

$$T_{\mathrm{JS}}(x_1, \dots, x_n) := (x_1, \dots, x_n)\frac{(\|x\|^2 - \sigma^2(n+2))_+}{\|x\|^2}$$

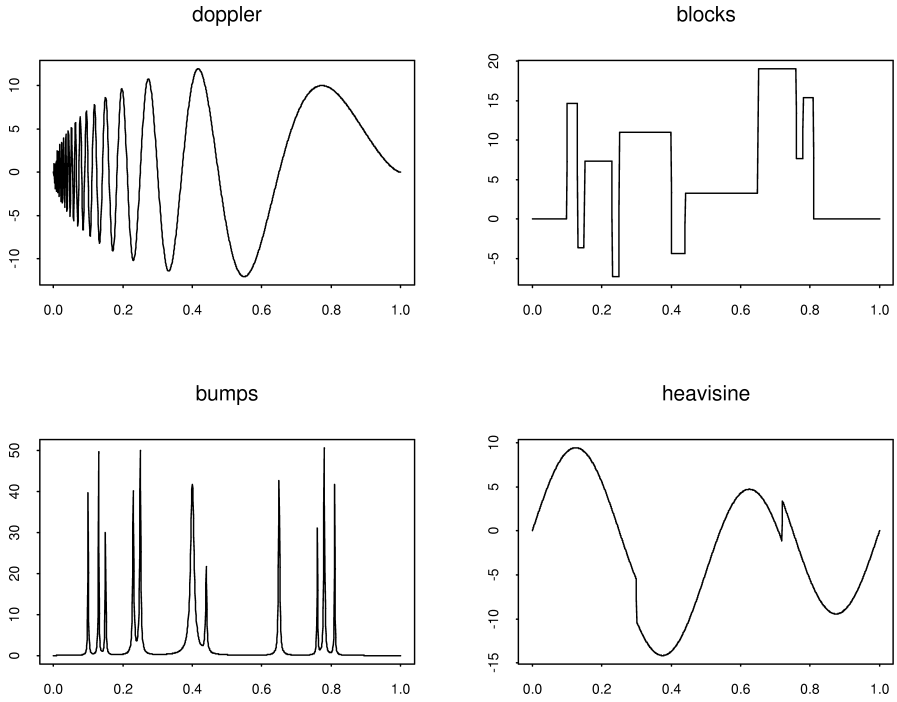

FIG. 4. *The four signals.*

is the James–Stein estimator. With this result, assume we are given a signal of length n which is contaminated by i.i.d. standard normal noise. Then the risk of any estimator which shrinks linearly each level of the wavelet transform of the data by a fixed amount is larger than the risk of the James–Stein estimator applied levelwise minus $2\log_2 n$ (we have $\log_2 n$ levels). Note that this property is independent of the signal itself. The situation might change if the linear shrinkage coefficients are chosen to depend on the noisy wavelet transform; this is what T_{JS} does.

A comparison of the performance of the different thresholds and of the performance of soft thresholding for non-Gaussian noise is of interest. To do this, a small Monte Carlo study was performed. The target signals are depicted in Figure 4; they were introduced by Donoho and Johnstone [13].

The simulation was performed with S+ from StatSci and the software package wavethresh for S+ from Guy Nason. As wavelet basis, the least asymmetric wavelets of Daubechies, with a filter length of 16, were chosen (see [8]). The lengths of the signal are 512 and 8192, while the noise is, respectively, normal and Student with four degrees of freedom, scaled to have variance 1. The density of this Student distribution decays like $1/|x|^5$, so its tails are quite heavy. The thresholds used are the optimal thresholds of Theorem 2.1 for normal noise and $n = 512$ and $n = 8192$, respectively (thresholding is done in the wavelet domain and so for Student noise, the central limit theorem dictates our choice

TABLE 1
*Optimal Gaussian thresholds**

	Signal length = 512						Signal length = 8192					
	Gaussian noise			Student noise			Gaussian noise			Student noise		
Signal	**0**	**3**	**5**	**0**	**3**	**5**	**0**	**3**	**5**	**0**	**3**	**5**
Doppler	0.45	0.40	0.35	0.51	0.49	0.39	0.074	0.07	0.06	0.099	0.095	0.094
Blocks	0.98	0.93	0.77	0.97	1.03	0.80	0.22	0.22	0.20	0.24	0.24	0.24
Bumps	1.11	1.12	1.02	1.17	1.17	1.07	0.21	0.21	0.19	0.24	0.23	0.23
Heavisine	0.24	0.18	0.15	0.34	0.22	0.21	0.046	0.043	0.034	0.074	0.064	0.068

*Average square errors for 100 runs.

of normal threshold). For each combination of noise, signal and signal length, different thresholding methods were applied: all levels are thresholded; the three coarsest levels are not thresholded and the five coarsest levels are not thresholded, denoted, respectively, by 0, 3 and 5 in Table 1. The numbers in the table are the averages of the square errors for 100 runs, divided by the length of the signal, that is, $\frac{1}{100}\sum_{k=1}^{100}\|\widehat{\theta}(\theta+e_k)-\theta\|_{2,n}^2$, where $\widehat{\theta}$ is the estimator (soft thresholding, James–Stein,...) and $\|\cdot\|_{2,n}$ the corresponding normalized (by n) Euclidean norm of $\theta=(\theta_1\ldots,\theta_n)$. Clearly, for a small sample size the levels which are not thresholded influence the performance of the estimator (as pointed out by a referee, this is much less an issue with hard thresholding).

Additionally, the estimator of Theorem 6.1 was used, but the coefficients of one level were thresholded with the largest threshold for that level of the original estimator; that is, the level j was thresholded with $\widetilde{\lambda}_{2^j+1}$ of Theorem 6.1.

For comparison the James–Stein estimator was applied levelwise to the wavelet coefficients, the five coarsest levels being treated as one level. This estimator tries to shrink the values with an estimate of the best linear shrinkage coefficient.

The main conclusion of this small study is that the performances of the estimator which does not threshold the five coarsest levels and of the estimator of

TABLE 2
Thresholds of Theorem 6.1*

	Gaussian noise		Student noise	
Signal	**512**	**8192**	**512**	**8192**
Doppler	0.39	0.046	0.49	0.095
Blocks	0.91	0.18	1.07	0.26
Bumps	1.16	0.16	1.20	0.25
Heavisine	0.17	0.028	0.23	0.070

*Average square errors for 100 runs.

TABLE 3
*James–Stein estimator**

	Gaussian noise		Student noise	
Signal	**512**	**8192**	**512**	**8192**
Doppler	0.55	0.077	0.58	0.078
Blocks	0.75	0.26	0.74	0.26
Bumps	0.91	0.14	0.92	0.14
Heavisine	0.18	0.040	0.19	0.043

*Average square errors for 100 runs.

Theorem 6.1 are comparable. Surprisingly good is the performance of the James–Stein estimator; it is also more robust if the noise is not normal.

To finish this study, the reader is referred to www.math.gatech.edu/∼houdre/ where an extensive set of simulations is presented. For various classes of noise, the visual appearance of the denoising method can also be evaluated there.

Acknowledgments. Both authors warmly thank L. Rüschendorf for making this collaboration possible and the referees for their very valuable comments which helped to improve the readability of the paper. The second author also thanks D. Donoho and I. Johnstone for discussions and support named, et, en dernier mais pas des moindres, Y. Meyer pour ses encouragements constants au cours des années.

REFERENCES

[1] ANTONIADIS, A. (1996). Smoothing noisy data with tapered coiflets series. *Scand. J. Statist.* **23** 313–330.

[2] AVERKAMP, R. and HOUDRÉ, C. (1999). Wavelet thresholding for non(necessarily) Gaussian noise: A preliminary report. In *Spline Functions and the Theory of Wavelets* (Montréal, PQ, 1996). *CRM Proc. Lecture Notes* **18** 347–354. AMS, Providence, RI.

[3] AVERKAMP, R. and HOUDRÉ, C. (1999). Wavelet thresholding for non(necessarily) Gaussian noise: Functionality. Preprint. Available at http://www.math.gatech.edu/∼houdre/

[4] BAKIROV, N. K. (1989). Extrema of the distributions of quadratic forms of Gaussian variables. *Theory Probab. Appl.* **34** 207–215.

[5] BICKEL, P. (1983). Minimax estimation of the mean of a normal distribution subject to doing well at a point. In *Recent Advances in Statistics* (M. H. Rizvi, J. S. Rustagi and D. Siegmund, eds.) 511–528. Academic Press, New York.

[6] BRUCE, A. and GAO, H. Y. (1995). Wave shrink with semisoft shrinkage. Statist. Sci. Research Report 39.

[7] CHAMBOLLE, A., DEVORE, R. A., LEE, N.-Y. and LUCIER, B. J. (1998). Nonlinear wavelet image processing: Variational problems, compression and noise removal through wavelet shrinkage. *IEEE Trans. Image Process.* **7** 319–335.

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

[9] DELYON, B. and JUDITSKY, A. (1995). Estimating wavelet coefficients. *Wavelets and Statistics. Lecture Notes in Statist.* **103** 151–168. Springer, Berlin.

[10] DELYON, B. and JUDITSKY, A. (1996). On minimax wavelet estimators. *Appl. Comput. Harmon. Anal.* **3** 215–228.

[11] DEVORE, R. A. and LUCIER, B. (1992). Fast wavelet techniques for near-optimal image processing. In *1992 IEEE Military Communications Conference* **3** 1129–1135. IEEE, New York.

[12] DONOHO, D. L. (1995). De-noising by soft-thresholding. *IEEE Trans. Inform. Theory* **41** 613–627.

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

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

[15] DONOHO, D. L. and JOHNSTONE, I. M. (1996). Neo-classical minimax problems, thresholding and adaptive function estimation. *Bernoulli* **2** 39–62.

[16] DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24** 508–539.

[17] EFRON, B. and MORRIS, C. (1971). Limiting the risk of Bayes and empirical Bayes estimators. I. The Bayes case. *J. Amer. Statist. Assoc.* **66** 807–815.

[18] FELLER, W. (1966). *An Introduction to Probability Theory and Its Applications* **2**. Wiley, New York.

[19] GAO, H. Y. (1993). Wavelet estimation of spectral densities in time series analysis. Ph.D. dissertation, Univ. California, Berkeley.

[20] GAO, H. Y. and BRUCE, A. G. (1997). WaveShrink with firm shrinkage. *Statist. Sinica* **7** 855–874.

[21] HÄRDLE, W., KERKYACHARIAN, G., PICARD, D. and TSYBAKOV, A. (1998). *Wavelets, Approximation and Statistical Applications. Lecture Notes in Statist.* **129**. Springer, Berlin.

[22] JOHNSTONE, I. M. and SILVERMAN, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *J. Roy. Statist. Soc. Ser. B* **59** 319–351.

[23] KERKYACHARIAN, G. and PICARD, D. (1992). Estimation de densité par méthode de noyau et d'ondelettes: Les liens entre la géometrie du noyau et les contraintes de régularité. *C. R. Acad. Sci. Paris Sér. I* **315** 79–84.

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

[25] KERKYACHARIAN, G. and PICARD, D. (1993). Density estimation by kernel and wavelets methods: Optimality of Besov spaces. *Statist. Probab. Lett.* **18** 327–336.

[26] KOLACZYK, E. D. (1997). Nonparametric estimation of gamma-ray burst intensities using Haar wavelets. *Astrophysical J.* **483** 340–349.

[27] KOLACZYK, E. D. (1999). Wavelet shrinkage estimation of certain Poisson intensity signals using corrected thresholds. *Statist. Sinica* **9** 119–136.

[28] KOVAC, A. and SILVERMAN, B. W. (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *J. Amer. Statist. Assoc.* **95** 172–183.

[29] LEDOUX, M. AND TALAGRAND, M. (1991). *Probability in Banach Spaces*. Springer, Berlin.

[30] MARSHALL, A. W. and OLKIN, I. (1979). *Inequalities*: *Theory of Majorization and Its Applications*. Academic Press, New York.

[31] MEYER, Y. (1990). *Ondelettes et Opérateurs* **1**. *Ondelettes*. Paris, Hermann.

[32] MEYER, Y. (1993). *Wavelets, Algorithms and Applications*. SIAM, Philadelphia.

[33] NEUMANN, M. H. and SPOKOINY, V. G. (1995). On the efficiency of wavelet estimators under arbitrary error distributions. *Math. Methods Statist.* **4** 137–166.

[34] PETROV, V. V. (1995). *Limit Theorems of Probability Theory*. Clarendon Press, Oxford.

[35] SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.
[36] STOUT, W. (1974). *Almost Sure Convergence*. Academic Press, New York.
[37] STRASSER, H. (1985). *Mathematical Theory of Statistics*. de Gruyter, Berlin.
[38] VIDAKOVIC, B. (1999). *Statistical Modeling by Wavelets*. Wiley, New York.
[39] WANG, Y. (1996). Function estimation via wavelet shrinkage for long-memory data. *Ann. Statist.* **24** 466–484.

INSTITUT FÜR MATHEMATISCHE STOCHASTIK
FREIBURG UNIVERSITY
79104 FREIBURG
GERMANY

LABORATOIRE D'ANALYSE
ET DE MATHÉMATIQUES APPLIQUÉES
CNRS UMR 8050
UNIVERSITÉ PARIS XII
94010 CRÉTEIL CEDEX
FRANCE
AND
SCHOOL OF MATHEMATICS
GEORGIA INSTITUTE OF TECHNOLOGY
ATLANTA, GEORGIA 30332
E-MAIL: houdre@math.gatech.edu