Asymptotic behavior of the minimum mean squared error threshold for noisy wavelet coefficients of piecewise smooth signals

Maarten Jansen
Adhemar Bultheel

Report TW 294, October 1999

Katholieke Universiteit Leuven
Department of Computer Science
Celestijnenlaan 200A – B-3001 Heverlee (Belgium)

Asymptotic behavior of the minimum mean squared error threshold for noisy wavelet coefficients of piecewise smooth signals

Maarten Jansen
Adhemar Bultheel

Report TW 294, October 1999

Department of Computer Science, K.U.Leuven

Abstract

This paper investigates the minimum risk threshold for wavelet coefficients with additive, homoscedastic, Gaussian noise, and for a soft-thresholding scheme. We start from N samples from a signal on a continuous time axis. For piecewise smooth signals, and for $N \to \infty$, this threshold behaves as $C\sqrt{2 \log N}\sigma$, where σ is the noise standard deviation. The paper contains an original proof for this asymptotic behavior as well as an intuitive explanation. This behavior is necessary to prove the asymptotic optimality of a generalized cross validation procedure in estimating the minimum risk threshold.

Keywords : Noise reduction, wavelet, thresholding, mean square error, risk, asymptotic.
AMS(MOS) Classification : 41A30, 93E14, 60G35, 62G07, 65D10, 62J02

Asymptotic behavior of the minimum mean squared error threshold for noisy wavelet coefficients of piecewise smooth signals

Maarten Jansen, Adhemar Bultheel

October 4, 1999

Abstract

This paper investigates the minimum risk threshold for wavelet coefficients with additive, homoscedastic, Gaussian noise, and for a soft-thresholding scheme. We start from N samples from a signal on a continuous time axis. For piecewise smooth signals, and for $N \to \infty$, this threshold behaves as $C\sqrt{2 \log N}\sigma$, where σ is the noise standard deviation. The paper contains an original proof for this asymptotic behavior as well as an intuitive explanation. This behavior is necessary to prove the asymptotic optimality of a generalized cross validation procedure in estimating the minimum risk threshold.

1 Introduction

A wavelet transform exploits the correlations between adjacent samples in a digital signal, to obtain a sparse data representation. This principle is also the basis for the popular wavelet threshold methods to reduce noise in signals: small wavelet coefficients are assumed to be dominated by noise and carry little information. Replacing these coefficients by zero eliminates a major part of the noise without affecting the signal too much. This paper investigates the mean squared error (MSE) and its expected value, called the risk function, as a criterion for selecting an optimal soft threshold for wavelet coefficients of piecewise smooth functions.

In applications, we never know the untouched coefficients, and we cannot possibly compute or minimize the exact MSE function. One possible procedure to estimate the minimum MSE threshold is based on generalized cross validation [10]. To prove that this method is asymptotically optimal, we need to know how the minimum risk threshold itself behaves if the number of samples tends to infinity. This motivates the study of this asymptotic expression.

We start from the classical additive model of a signal $\boldsymbol{f}$ corrupted with noise $\boldsymbol{\eta}$:

$$\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\eta}. \tag{1}$$

The vector $\boldsymbol{y}$ represents the input signal. The noise is a vector of random variables, while the untouched values $\boldsymbol{f}$ form a purely deterministic signal. let N be the length of these vectors. The noise is assumed to be zero mean and homoscedastic or second order stationary. This means that all noise components η_i have the same standard deviation. We work with normal (Gaussian) density functions.

A wavelet transform $\tilde{W}$ is linear, and so leaves this additivity unchanged:

$$\boldsymbol{w} = \boldsymbol{v} + \boldsymbol{\omega}, \tag{2}$$

where $\boldsymbol{v}$ is the vector of uncorrupted (untouched, noise-free) wavelet coefficients, $\boldsymbol{\omega}$ contains the wavelet transform of the noise and $\boldsymbol{w}$ are the wavelet coefficients of the observed signal. Normality is also preserved, but homoscedasticy is lost if the input noise is correlated or the transform is not orthogonal. For the purpose of this text, we need homoscedasticy, and so we assume orthogonal transforms and uncorrelated input noise. Relaxation of these conditions is possible through level-dependent thresholds [11, 9].

In the first section, we introduce the mean square error as a function of the threshold value, and examine its typical shape. Next, we focus on the threshold that minimizes this objective function. We try to understand how it behaves asymptotically, i.e. if the number of data N tends to infinity. We first deal with the piecewise polynomial case, because this causes many coefficients to be exactly zero, and this facilitates the analysis. Next, we proceed to general piecewise smooth functions. The mean square error has also been analyzed in other papers [1, 6, 7].

2 Mean square error and Risk function

2.1 Definitions

A threshold can be seen as a smoothing parameter: it controls the compromise between goodness of fit and smoothness of approximation. In this context, smoothness should be interpreted as sparsity: we try to find a sparse data set, close to the noisy input.

The ultimate objective is of course an approximation of the noise-free data. While balancing between closeness of fit and sparsity, the *best* compromise minimizes the error of the result as compared with these unknown, uncorrupted data.

If $\boldsymbol{y_\lambda}$ is the output of the threshold algorithm with some threshold value λ and $\boldsymbol{f}$ is the vector of untouched data, the remaining noise on this result equals $\boldsymbol{\eta}_{\boldsymbol{\lambda}} = \boldsymbol{y}_{\boldsymbol{\lambda}} - \boldsymbol{f}$, and the mean squared error (MSE) is then defined as:

$$R(\lambda) = \mathrm{MSE}(\lambda) = \frac{1}{N}\|\boldsymbol{\eta}_{\boldsymbol{\lambda}}\|^2. \tag{3}$$

As the notation indicates, the MSE, R(λ), is a function of the threshold value λ. It is also a random variable, because it depends on the noise. The expected value of this error is called the *risk*-function.

The main challenge with this MSE as an objective function is the fact that in real applications, it can never be computed exactly: its definition uses the value of the exact, unknown data $\boldsymbol{f}$. In practical situations, this MSE has to be estimated [5, 10, 13].

A common definition of signal-to-noise ratio (SNR) is based on this notion of MSE:

$$\mathrm{SNR}(\lambda) = 10 \cdot \log_{10} \frac{\|\boldsymbol{f}\|^2}{\|\boldsymbol{\eta}_{\boldsymbol{\lambda}}\|^2} = 10 \cdot \log_{10} \frac{\|\boldsymbol{f}\|^2/N}{R(\lambda)}. \tag{4}$$

An alternative is the peak signal-to-noise ratio, which is equal to the previous one, up to constant, depending on the uncorrupted data:

$$\mathrm{PSNR}(\lambda) = 10 \cdot \log_{10} \frac{(\max \boldsymbol{f})^2/N}{R(\lambda)} = \mathrm{SNR}(\lambda) + 10 \cdot \log_{10} \frac{(\max \boldsymbol{f})^2}{\|\boldsymbol{f}\|^2}. \tag{5}$$

An orthogonal wavelet transform $\tilde{W}$ preserves the ℓ_2-norm, and so:

$$R(\lambda) = \frac{1}{N}\|\boldsymbol{\omega}_{\boldsymbol{\lambda}}\|^2,$$

where $\boldsymbol{\omega}_{\boldsymbol{\lambda}} = \boldsymbol{w}_{\boldsymbol{\lambda}} - \boldsymbol{v} = \tilde{W}(\boldsymbol{y}_{\boldsymbol{\lambda}} - \boldsymbol{f})$. From now on, we do all our computations in the wavelet domain. If the transform is biorthogonal, there is no exact equivalence with the data domain. Nevertheless, computation and minimization in terms of wavelet coefficients seems to give satisfactory results, and several reasons could explain this: Riesz-bounds guarantee a nearly equivalent norm. Moreover, since MSE does not correspond exactly to a human perception of quality, the question arises whether MSE in the original data domain is always a better measure than MSE in the wavelet domain. In image processing applications, for instance, we view the image in the pixel domain, but we do not look at an image as a matrix of pixels. Since our visual system seems to work on a multiscale basis, a norm based on a multiresolution decomposition might be a better expression of visual quality. Further illustrations show that there is no need for expressing norms in the original data domain. This preserves us from applying an inverse wavelet transform every time we want to evaluate the quality of a result. An inverse wavelet transform is only necessary to compute the eventual output of the algorithm.

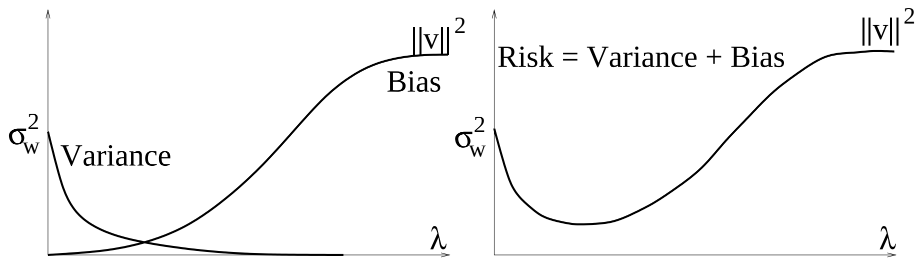

Figure 1: Typical behavior of bias and variance as a function of the threshold value. Thresholding introduces bias, but reduces variance. The best compromise minimizes the risk.

2.2 Variance and bias

The input wavelet coefficients are unbiased estimates of the noise-free coefficients:

$$\mathrm{E}\boldsymbol{w} = \mathrm{E}\tilde{W}\boldsymbol{y} = \tilde{W}\mathrm{E}\boldsymbol{y} = \tilde{W}\boldsymbol{f} = \boldsymbol{v}$$

but the variance of this "estimation" is too high. Replacing the smallest coefficients with zero reduces the variance, at the cost of an increasing bias:

$$\mathrm{bias}(\lambda) = \frac{1}{N}\|\mathrm{E}\boldsymbol{w}_{\boldsymbol{\lambda}} - \boldsymbol{v}\|^2 \quad (6)$$

$$\mathrm{variance}(\lambda) = \frac{1}{N}\mathrm{E}\|\boldsymbol{w}_{\boldsymbol{\lambda}} - \mathrm{E}\boldsymbol{w}_{\boldsymbol{\lambda}}\|^2. \quad (7)$$

Then it holds that:

$$\begin{aligned} \mathrm{E}R(\lambda) &= \frac{1}{N}\|\mathrm{E}\boldsymbol{w}_{\boldsymbol{\lambda}} - \boldsymbol{v}\|^2 + \frac{1}{N}\mathrm{E}\|\boldsymbol{w}_{\boldsymbol{\lambda}} - \mathrm{E}\boldsymbol{w}_{\boldsymbol{\lambda}}\|^2 \quad (8) \\ \text{Risk} &= \qquad \text{bias} \qquad + \qquad \text{variance} \end{aligned}$$

The most reliable method to remove *all* noise is just removing everything:

$$\lim_{\lambda\to\infty} \mathrm{variance}(\lambda) = 0.$$

If all coefficients are removed, there is no variance anymore, all the noise has gone, but so has the signal: the bias equals the total energy of the noise-free input:

$$\lim_{\lambda\to\infty} \mathrm{bias}(\lambda) = \|\boldsymbol{v}\|^2$$

Figure 1 shows a typical behavior of these functions. The minimum risk threshold is the best compromise (in ℓ_2) between variance and bias.

3 The risk contribution of each coefficient (Gaussian noise)

This section puts some elementary calculations together. The results are necessary for the next sections. From now on, we assume that the input noise is Gaussian and we call:

$$\begin{aligned}\phi(\omega) &= \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{\omega^2/2\sigma^2},\\ \Phi(\omega) &= \int_{-\infty}^{\omega}\phi(u)du.\end{aligned}$$

Every classical, linear wavelet transform preserves the normality of a density. If the input noise is not Gaussian, the density of the wavelet coefficients, if at all computable in practice, would depend on the type of wavelets being used.

Some of the following results also appear in different papers like [4]. A first lemma gives an expression for the bias of one coefficient $w = v + \omega$ (the notation omits the index of the coefficient).

Lemma 1

$$\mathrm{E}\omega_\lambda = \sigma^2[\phi(\lambda-v)-\phi(\lambda+v)]+\lambda[\Phi(\lambda-v)-\Phi(\lambda+v)]+v[1-\Phi(\lambda-v)-\Phi(\lambda+v)]. \tag{9}$$

The proof is by simple calculations, using the fact that a Gaussian distribution satisfies the following differential equation:

$$\omega\phi(\omega) = -\sigma^2\phi'(\omega). \tag{10}$$

We denote by

$$r(v,\lambda) = \mathrm{E}(w_\lambda - v)^2 \tag{11}$$

the contribution of coefficient w to the total risk function. Using Equation (10) and partial integration leads to

$$\int \omega^2\phi(\omega)d\omega = -\omega\sigma^2\phi(\omega) + \sigma^2\int\phi(\omega)d\omega,$$

which allows to conclude, after some calculation, that:

Lemma 2

$$\begin{aligned}r(v,\lambda) &= \left[2(\sigma^2+\lambda^2)-v^2\right] + [\Phi(\lambda-v)+\Phi(\lambda+v)]\,(v^2-\sigma^2-\lambda^2)\\ &\quad -\sigma^2\left[(\lambda-v)\phi(\lambda+v)+(\lambda+v)\phi(\lambda-v)\right].\end{aligned} \tag{12}$$

Plots of this contribution as a function of the threshold λ for various values of v show that coefficients with little information ($v \approx 0$) are best served with large thresholds, whereas important coefficients (v large) prefer little thresholding. The overall optimal threshold is the best compromise between these two.

To find the minima, we compute the derivative. Again an trivial computation leads to:

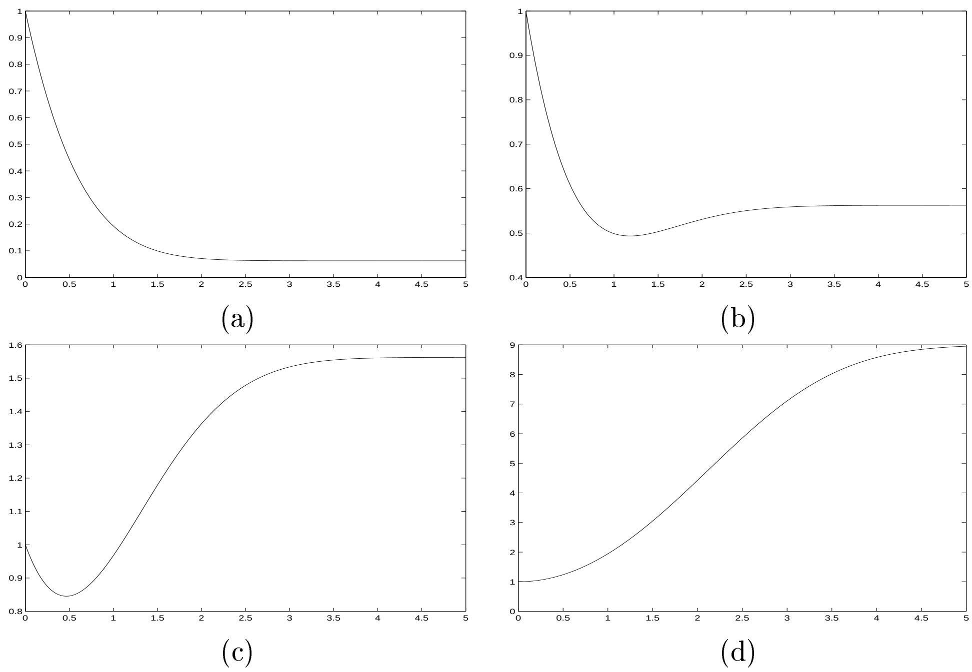

Figure 2: Contribution of individual coefficients to the total risk as a function of the threshold value. Small values of v (a) prefer large thresholds, because bias is small. Large values of v (d) would cause considerable bias if the threshold gets large. (b) and (c) are two typical, intermediate cases.

Lemma 3

$$\frac{\partial r}{\partial \lambda}(v,\lambda) = 2\lambda\,[1+\Phi(-v-\lambda)-\Phi(-v+\lambda)]-2\sigma^2\,[\phi(v+\lambda)+\phi(-v+\lambda)]. \quad (13)$$

The proof uses the fact that

$$\frac{\partial}{\partial \lambda}\mathrm{E}(w_\lambda - v)^2 = \mathrm{E}\frac{\partial}{\partial \lambda}(w_\lambda - v)^2.$$

An important case is that of a coefficient without any information. It turns out that if $v = 0$, the derivative $\frac{\partial r}{\partial \lambda}$ is always negative (see Figure 2, upper left). If $\lambda \to \infty$, the derivative approaches zero, but it remains negative. This means that the optimal threshold for this zero coefficient equals infinity. This is confirmed by the following asymptotic behavior:

Lemma 4

$$\frac{\partial r}{\partial \lambda}(0,\lambda) \sim -4\sigma^4\frac{\phi(\lambda)}{\lambda^2}. \quad (14)$$

Proof:
From the previous lemma, we see that:

$$\frac{\partial r}{\partial \lambda}(0,\lambda) = 4\lambda[1-\Phi(\lambda)] - 4\sigma^2\phi(\lambda).$$

Three times De L'Hôpital's rule shows that:

$$\lim_{\lambda\to\infty}\frac{\lambda[1-\Phi(\lambda)]-\sigma^2\phi(\lambda)}{\frac{\phi(\lambda)}{\lambda^2}} = \sigma^4.$$

□

To get an idea of how $\frac{\partial r}{\partial \lambda}$ behaves more generally, we compute the derivative of this expression with respect to the uncorrupted coefficient value v:

Lemma 5

$$\frac{\partial}{\partial v}\left(\frac{\partial r}{\partial \lambda}(v,\lambda)\right) = 2v\,[\phi(v+\lambda)+\phi(-v+\lambda)] \quad \begin{cases} \geq 0 & \textit{if } v \geq 0, \\ \leq 0 & \textit{if } v \leq 0. \end{cases}$$

Consequently, for a given threshold, $\frac{\partial r}{\partial \lambda}$ has a minimum for $v = 0$. Figure 3 shows $\frac{\partial r}{\partial \lambda}$ as a function of v for two different values of λ. The plot on the left-hand side corresponds to a threshold $\lambda = 0.5\,\sigma$. This threshold is too small for all coefficients smaller than approximately $1.1\,\sigma$. If we choose $\lambda = 2\sigma$,

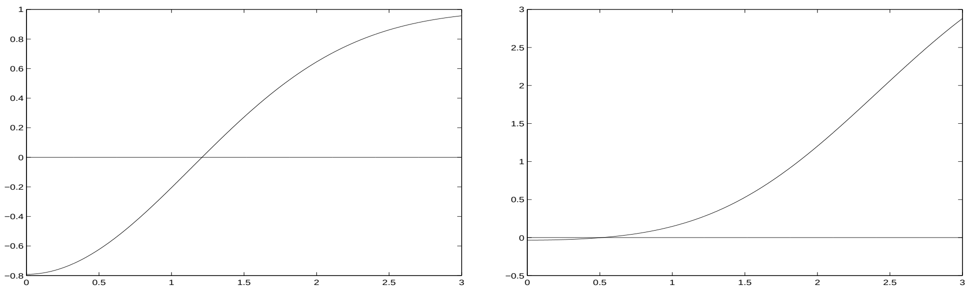

Figure 3: Derivative of the risk in a given coefficient with respect to the threshold value as a function of the noise-free coefficient value. Left: $\lambda = 0.5\,\sigma$. For coefficients approximately below $1.1\,\sigma$, this threshold is too low. Right: $\lambda = 2\sigma$. This threshold is too large for coefficients above $0.5\,\sigma$. The distribution of noise-free coefficients determines the optimal threshold.

only coefficients below $0.5\,\sigma$ find this value too small. The value of the optimal threshold depends on how the noise-free coefficients are distributed: the sparser the representation is, the larger the optimal threshold will be. Indeed, if the proportion of small coefficients increases, the threshold should be large, because all these small coefficients prefer large thresholds. The next sections try to find an asymptotic behavior for this optimal threshold. We assume that generating more samples from a given signal on a continuous line, introduces more redundancy in the information. This causes more sparsity in the wavelet representation. We expect that the optimal threshold increases as the number of samples N grows.

4 The asymptotic behavior of the minimum risk threshold for piecewise polynomials

4.1 Motivation

It can be shown [10] that, in minimum risk sense, the GCV-method asymptotically yields the optimal threshold. This property motivates the use of GCV in a threshold assessment procedure. For the proof of this asymptotic optimality, we need to know how the optimal threshold itself behaves if the number of samples $N \to \infty$. This section assumes that the samples come from a piecewise polynomial on $[0, 1]$:

$$y_i = f(\Delta t \cdot i) + \eta_i,$$

where $f(t)$ is a piecewise polynomial and $t \in [0, 1]$. No real signal is of course a perfect piecewise polynomial, but typical signals are piecewise smooth. Section 5 investigates how the threshold behaves in this more general case.

4.2 Asymptotic equivalence

Before studying the asymptotics of the minimum risk threshold, we recall the definition of asymptotic equivalence:

Definition 1 *Two functions $f(x)$ and $g(x)$ are said to be asymptotic equivalent for $x \to \infty$, i.e. $f(x) \sim g(x)$, if and only if*

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = 1.$$

The study of the asymptotics of the minimum risk threshold uses a couple of properties of this notion:

Lemma 6 *Let $x \to \infty$. If $f(x) \sim g(x)$, we have:*

1. *$f(x)h(x) \sim g(x)h(x)$.*
2. *If $f(x) \sim h(x)$, then $h(x) \sim g(x)$.*
3. *If $h(x) = o(f(x))$, then $f(x) \pm h(x) \sim g(x)$.*
4. *If*
$$\lim_{x \to \infty} f(x) \neq 1 \neq \lim_{x \to \infty} g(x),$$
and both functions are differentiable, then
$$\log f(x) \sim \log g(x).$$

Proof:

1. Trivialiter
2. $$\lim_{x \to \infty} \frac{h(x)}{g(x)} = \lim_{x \to \infty} \frac{h(x)}{f(x)} \frac{f(x)}{g(x)},$$
and we may split the limit of this product, since both factors have a limit.
3. This is actually a special case of the previous statement.
$$\lim_{x \to \infty} \frac{f(x) \pm h(x)}{g(x)} = \lim_{x \to \infty} \frac{f(x)}{g(x)} \left(1 + \frac{h(x)}{f(x)}\right) = \lim_{x \to \infty} \frac{f(x)}{g(x)} \cdot \lim_{x \to \infty} \left(1 + \frac{h(x)}{f(x)}\right)$$
4. For $\lim_{x \to \infty} f(x) = \infty$, we use De L'Hôpital's rule in two directions: we use De L'Hôpital's rule:
$$\begin{aligned} \lim_{x \to \infty} \frac{\log f(x)}{\log g(x)} &= \lim_{x \to \infty} \frac{f'(x)/f(x)}{g'(x)/g(x)} = \lim_{x \to \infty} \frac{f'(x)}{g'(x)} \lim_{x \to \infty} \frac{g(x)}{f(x)} \\ &= \lim_{x \to \infty} \frac{f(x)}{g(x)} \lim_{x \to \infty} \frac{g(x)}{f(x)} = 1. \end{aligned}$$

For a finite limit, we do not need the differentiability:

$$\lim_{x\to\infty} \frac{\log f(x)}{\log g(x)} = \frac{\lim\limits_{x\to\infty} \log f(x)}{\lim\limits_{x\to\infty} \log g(x)} = \frac{\log\left(\lim\limits_{x\to\infty} f(x)\right)}{\log\left(\lim\limits_{x\to\infty} g(x)\right)} = 1$$

□

We remark that the inverse implication of the last statement definitely does not hold: if $\log f(x) \sim \log g(x)$, $f(x)$ and $g(x)$ may be not asymptotically equivalent. For instance, if $\log f(x) = x + \sqrt{x}$ and $\log g(x) = x$, then

$$\frac{f(x)}{g(x)} = \mathrm{e}^{\sqrt{x}} \not\to 1.$$

4.3 The asymptotic behavior

For the piecewise polynomial case, we assume that the wavelet analysis has more vanishing moments than the highest degree of the polynomials. As a consequence, wavelet coefficients are zero if they do not correspond to a basis function which interferes with a singularity. We assume that the number of singularities is finite on $[0, 1]$.

We then have the following theorem for the asymptotic behavior of the minimum of $\mathrm{E}R(\lambda) = \mathrm{E}\|\boldsymbol{y}_{\boldsymbol{\lambda}} - \boldsymbol{f}\|^2$:

Theorem 1 *If λ^* minimizes $\mathrm{E}R(\lambda)$, then for $N \to \infty$:*

$$\lambda^* \sim \sqrt{2 \log N}\sigma \tag{15}$$

Proof:
We suppose that the wavelet transform is orthogonal, so the problem model in the wavelet domain is the same as in the input (time or space) domain:

$$\boldsymbol{w} = \boldsymbol{v} + \boldsymbol{\omega},$$

where $\boldsymbol{\omega}$ is i.i.d. noise with variance σ^2. A wavelet coefficient v_i or w_i corresponds to a basis function $\psi_{j,k}$ at resolution level j and place k.

We call:

$$\begin{aligned} I_0 &= \{i = 1, \ldots, N | v_i = 0\} \\ I_1 &= \{i = 1, \ldots, N | v_i \neq 0\} \\ M_0 &= \# I_0 \\ M_1 &= \# I_1 \end{aligned}$$

M_0 and M_1 of course depend on N. Since $f(t)$ is a piecewise polynomial, at each level only a constant number of coefficients is not exactly zero. The total number of non-zero coefficients is proportional to the number of levels:

$$M_1 \sim \log N,$$

and so:

$$\frac{M_1}{N} \to 0.$$

Using the notation from the previous section, we may write:

$$\mathrm{E}R(\lambda) = \sum_{i=1}^{N} r(v_i, \lambda).$$

And so:

$$\mathrm{E}R'(\lambda) = M_0 \frac{\partial r}{\partial \lambda}(0, \lambda) + \sum_{i \in I_1} \frac{\partial r}{\partial \lambda}(v_i, \lambda).$$

Lemma 3 learns that $\frac{\partial r}{\partial \lambda}(0, \lambda) < 0$. Call $I_1' \subset I_1$ the indices of the non-zeros for which $\frac{\partial r}{\partial \lambda}(v_i, \lambda^*)$ is negative. These indices belong to the smaller coefficients. The indices of the large coefficients are in $I_1'' = I_1 \backslash I_1'$. We define

$$\begin{aligned} M_1' &= \# I_1' \\ M_1'' &= \# I_1''. \end{aligned}$$

We know that

$$v_i \approx \sqrt{N} \int_{\mathbb{R}} f(t) \psi_{j,k}(t) dt,$$

where the integral is a constant value which does not depend on N, so $v_i \sim \sqrt{N}$, and we are looking for a λ^* which does not increases faster. This means that M_1'' is a non-decreasing function of N: if $N \to \infty$, ever more coefficients are classified as large, since all non-zero coefficients grow at least as fast as the optimal threshold. On the other hand, $M_1'' \leq M_1$ does not increase too fast. We now write the equation for λ^*: $\mathrm{E}R'(\lambda^*) = 0$ or:

$$-M_0 \frac{\partial r}{\partial \lambda}(0, \lambda^*) - \sum_{i \in I_1'} \frac{\partial r}{\partial \lambda}(v_i, \lambda^*) = \sum_{i \in I_1''} \frac{\partial r}{\partial \lambda}(v_i, \lambda^*). \qquad (16)$$

We consider this equation as an equality of two functions of N, and let $N \to \infty$. Both sides of this equation have all positive terms. Lemma 5 says that

$$\forall i \in I_1' : \left| \frac{\partial r}{\partial \lambda}(v_i, \lambda^*) \right| \leq \left| \frac{\partial r}{\partial \lambda}(0, \lambda^*) \right|.$$

Moreover $M_1' \leq M_1 \sim \log N$, so $\frac{M_1'}{M_0} \sim \frac{\log N}{N} \to 0$. From this, we may conclude that the sum $\sum_{i \in I_1'}$ in Equation (16) can be neglected.

If v_i grows faster than λ^* for increasing N, the right-hand side behaves like $M_1''2\lambda^*$, as follows from letting $v \to \infty$ in Lemma 3. We have:

$$\frac{M_0}{M_1''} \sim \frac{2\lambda^*}{-\frac{\partial r}{\partial \lambda}(0,\lambda^*)}.$$

If $N \to \infty$, the left-hand side grows like $N/\log N \to \infty$. The right-hand side is an increasing function of λ^*: it is easy to verify (from the proof of Lemma 4) that

$$-\frac{\partial^2 r}{\partial \lambda^2}(0,\lambda) = -4\left[1 - \Phi(\lambda)\right] \leq 0,$$

so the denominator is a positive, decreasing function, while the numerator is positive and increasing. To make this right-hand side grow to infinity, we need $\lambda^* \to \infty$. Therefore, we can use Lemma 4 and get the following asymptotic equation:

$$\begin{aligned}
\frac{M_0}{M_1''} &\sim \frac{2\lambda^*\lambda^{*2}}{4\sigma^4\phi(\lambda^*)}, \\
\frac{\sigma^3 M_0}{M_1''} &\sim \sqrt{\frac{\pi}{2}}\, e^{\lambda^{*2}/2\sigma^2}\lambda^{*3}, \\
3\log\sigma + \log M_0 - \log M_1'' &\sim \frac{\lambda^{*2}}{2\sigma^2} + 3\log\lambda^* + \frac{1}{2}\log\frac{\pi}{2}.
\end{aligned}$$

The left-hand side depends on N, the right-hand side depends on λ^*. We keep the essential on both sides:

$$\begin{aligned}
\log M_0 &\sim \frac{\lambda^{*2}}{2\sigma^2}, \\
\log(N - M_1) &\sim \frac{\lambda^{*2}}{2\sigma^2}, \\
2\sigma^2 \log N &\sim \lambda^{*2}. \qquad \square
\end{aligned}$$

4.4 An example

Figure 4 shows the plot of a piecewise, linear polynomial. This function is sampled, and transformed into wavelet domain, using the orthogonal Daubechies wavelets with two vanishing moments. We then compute numerically the minimum of

$$\mathrm{E}R(\lambda)$$

for different sample rates, and $\sigma = 1$. These values are listed in Table 1 and plotted in Figure 5.

Both table and figure illustrate that indeed

$$\lambda^* \approx K + \sqrt{2\log N}\sigma = K + \sqrt{2\log 2}\sqrt{J},$$

where K is constant and $2^J = N$.

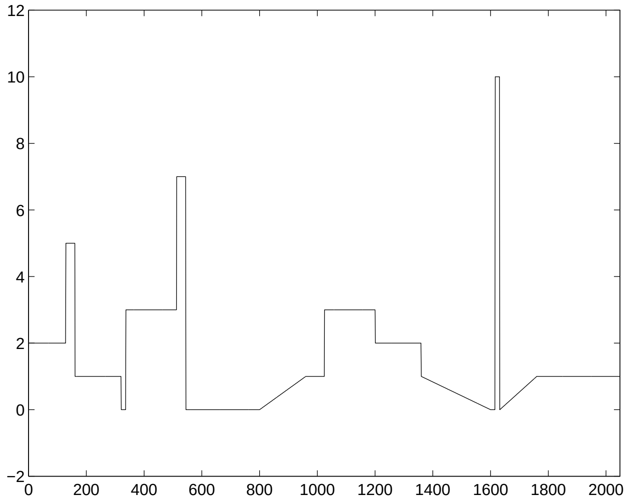

Figure 4: An example of a piecewise polynomial: in this case, all pieces are linear or constant.

J	$N = 2^J$	λ^*
7	128	1.00564887172677
8	256	1.12338708120546
9	512	1.25150560042977
10	1024	1.40212517387946
11	2048	1.55837014248141
12	4096	1.75818138547032
13	8192	1.94789562337191
14	16384	2.13051380127175
15	32768	2.30620858768226
16	65536	2.47521770450833
17	131072	2.63789870923179
18	262144	2.79627013633284
19	524288	2.95567796055755
20	1048576	3.11403413842011

Table 1: Minimum risk threshold for the piecewise polynomial in Figure 4 as a function of the number N of samples.

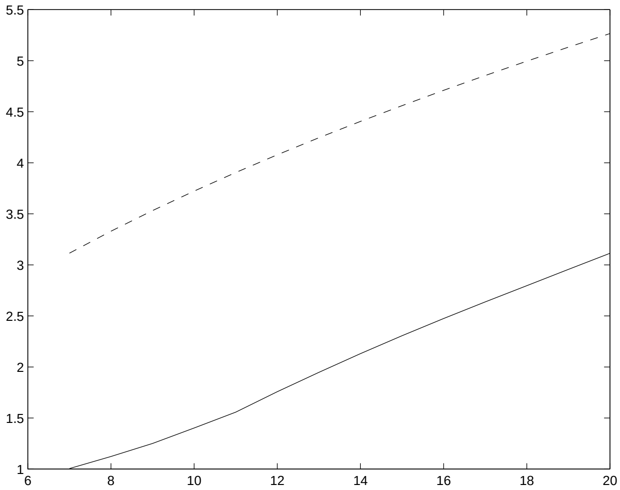

Figure 5: Plot of minimum risk threshold for the signal in Figure 4 as a function of the binary logarithm J of the number of samples (full line). De dashed line is a plot of the predicted equivalence: $\sqrt{2\log 2}\sqrt{J}$. The plot seems to confirm this asymptotic behavior.

4.5 Why does the threshold depend on the number of data points?

To engineers it might look strange that the optimal threshold depends on the number of data points. They object that the threshold should not change by putting two signals together?

First, this objection does not correspond to the philosophy behind this asymptotic analysis: we do not join two signals, but merely take more samples from one function on a given interval. Second, as Table 1 illustrates, we note that $\sqrt{2\log N}$ is only a very weak dependence.

And third, there is a comprehensive explanation for this behavior. Adding more samples enhances redundancy in the signal: there is less new information in new samples than there was in the first samples. In wavelet domain, this means that the number of important coefficients is hardly growing, and all information remains concentrated in a limited number of coefficients. If we suppose that the transform is normalized, the magnitude of these large coefficients should increase, since more samples mean a higher total energy (2-norm of the data vector) and this energy is preserved by the wavelet transform, while all nearly zero coefficients hardly take any of it. On the other hand, the noise variance in all coefficients remains σ^2 all the time. If the threshold would be independent of N, say $\lambda = k\sigma$, then the relative number of purely noise coefficients which passes the threshold would converge to $\mathrm{P}(|Z| > k)$, Z being a standard normal variable. So, the total number of noise coefficients would be proportional to N. Since the number of important coefficients is approximately a constant, the reconstruction would become noisier. Therefore, it is better to let the threshold increase slowly to catch all noise coefficients, while leaving the faster growing signal coefficients intact.

4.6 Universal Threshold

Of course, the formula of the asymptotic behavior of the minimum risk threshold does not tell everything about the actual, optimal threshold value. This actual value depends on all coefficients, while the asymptotic formula only depends on N and σ. We did not say that one should use this asymptotic formula as a real threshold. Nevertheless, the value

$$\lambda_{\text{UNIV}} = \sqrt{2\log N}\sigma \tag{17}$$

is well known in wavelet literature and it is used as a threshold value, not only as an asymptotic equivalence. This is the so-called *universal threshold.* This name reflects the idea that this threshold is "valid" for all signals with length N, provided that these signals are "sufficiently" smooth: it is a general threshold value. Donoho, Johnstone and collaborators have proven a lot of optimality properties for this choice [4, 2, 3, 5].

5 Beyond the piecewise polynomial case

Theorem 1 investigated the asymptotic behavior of the minimum risk threshold λ^* for piecewise polynomials. We would like to generalize this result to general piecewise smooth functions. The proof of the polynomial theorem introduced the idea that the optimal threshold is the best compromise between coefficients with a large uncorrupted value, for which this threshold is already beyond the optimum, $(\frac{\partial r}{\partial \lambda}(v_i, \lambda^*) > 0)$ and small coefficients, for which the optimal threshold could be larger $(\frac{\partial r}{\partial \lambda}(v_i, \lambda^*) < 0)$. This distinction implicitly divides the coefficients into two groups, but we did not compute the boundary v^* between them, because for piecewise polynomials, we can count on the important group of coefficients exactly equal to zero.

For general piecewise smooth functions, none of the coefficients is exactly zero, and therefore we want to have an idea for which and for how many coefficients a given threshold λ^* is too large or too small. This could give us an impression of the behavior of the optimal compromise.

5.1 For which coefficients is a given threshold too large/small?

From Lemma 3, we learn that:

$$\frac{\partial r}{\partial \lambda}(v, \lambda) = 0 \Leftrightarrow \frac{\sigma^2}{\lambda} = \frac{1 + \Phi(-v - \lambda) - \Phi(-v + \lambda)}{\phi(v + \lambda) + \phi(-v + \lambda)}. \tag{18}$$

We now consider this as an equation in v and look for a lower bound for its solution v^* as a function of λ. Lemma 5 says that for a fixed threshold value, $\frac{\partial r}{\partial \lambda}(v, \lambda)$ is a monotonically increasing function of v if $v > 0$, and this guarantees that Equation (18) has at most one solution.

Let $G(v, \lambda)$ be the right-hand side of Equation (18):

$$G(v, \lambda) = \frac{1 + \Phi(-v - \lambda) - \Phi(-v + \lambda)}{\phi(v + \lambda) + \phi(-v + \lambda)}.$$

It is trivial to see that

$$\lim_{v \to \infty} G(v, \lambda) = \infty.$$

If we find a value v^0 for which $G(v^0, \lambda) \leq \frac{\sigma^2}{\lambda}$, we may conclude that the solution v^* of (18) satisfies $v^* \geq v^0$.

We evaluate

$$G\left(\frac{\sigma^2}{\lambda}, \lambda\right) = \sigma \, \frac{2 - \Phi_1(u + 1/u) - \Phi_1(u - 1/u)}{\phi_1(u + 1/u) + \phi_1(u - 1/u)},$$

in which $u = \lambda/\sigma$ and ϕ_1, Φ_1 are standard normal density and distribution:

$$\phi_1(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}.$$

The next, technical section argues that:

$$u \, \frac{2 - \Phi_1(u + 1/u) - \Phi_1(u - 1/u)}{\phi_1(u + 1/u) + \phi_1(u - 1/u)} \leq 1 \qquad (19)$$

for all $u \geq 1.7815$, and so

$$G\left(\frac{\sigma^2}{\lambda}, \lambda\right) \leq \frac{\sigma^2}{\lambda}$$

if $\lambda \geq 1.7815\,\sigma$.

This allows us to formulate the following theorem:

Theorem 2 *If $\lambda \geq 1.7815\,\sigma$, and v^* satisfies*

$$\frac{\partial r}{\partial \lambda}(v^*, \lambda) = 0$$

then

$$v^* \geq \frac{\sigma^2}{\lambda}. \qquad (20)$$

Figure 6 shows a numerical computation of the curve $v^*(\lambda)$. It demonstrates that we have found a sharp lower bound.

In the upcoming analysis we need the fact that $\frac{\partial r}{\partial \lambda}(v, \lambda)$ is convex as a function of v for $|v| \leq \sigma^2/\lambda$. Therefore, we formulate an additional lemma:

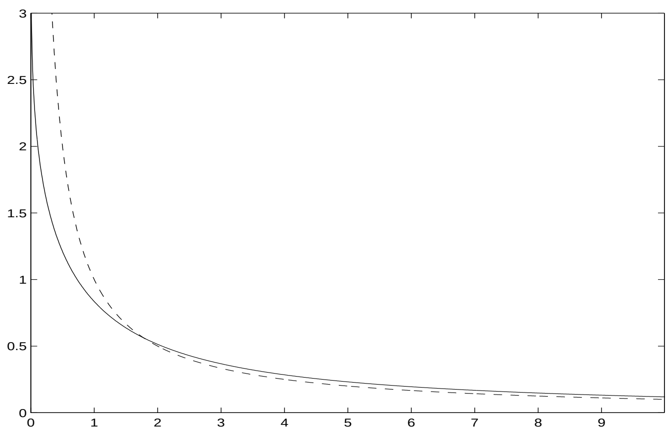

Figure 6: Full line: plot of v^* as a function of λ, where v^* satisfies $\frac{\partial r}{\partial \lambda}(v^*, \lambda) = 0$. Dashed line: plot of σ^2/λ. In this example we put $\sigma = 1$.

Lemma 7 *If $\lambda \geq \sigma$, we have for $v \leq \sigma^2/\lambda$:*

$$\frac{\partial^2}{\partial v^2}\left(\frac{\partial r}{\partial \lambda}\right) \geq 0. \tag{21}$$

Proof:
From Lemma 5 we compute

$$\frac{\partial^2}{\partial v^2}\left(\frac{\partial r}{\partial \lambda}\right) = 2\phi(v+\lambda)\left[1 - \frac{v(v+\lambda)}{\sigma^2}\right] + 2\phi(v-\lambda)\left[1 - \frac{v(v-\lambda)}{\sigma^2}\right]. \tag{22}$$

The factor

$$1 - \frac{v(v-\lambda)}{\sigma^2}$$

is positive on

$$\left[\frac{\lambda - \sqrt{\lambda^2 + 4\sigma^2}}{2}, \frac{\lambda + \sqrt{\lambda^2 + 4\sigma^2}}{2}\right],$$

which contains the interval $[0, \lambda]$.

Since $v \leq \sigma^2/\lambda$ and $\lambda \geq \sigma$ by assumption, we have $v \leq \lambda$, and so:

$$\begin{aligned} \frac{\partial^2}{\partial v^2}\left(\frac{\partial r}{\partial \lambda}\right) &\geq 2\phi(v+\lambda)\left[1 - \frac{v(v+\lambda)}{\sigma^2} + 1 - \frac{v(v-\lambda)}{\sigma^2}\right] \\ &= 4\phi(v+\lambda)\left[1 - \frac{v^2}{\sigma^2}\right]. \end{aligned}$$

From $v \leq \sigma^2/\lambda$ and $\lambda \geq \sigma$, it also follows that $v \leq \sigma$, and so we know that this expression is positive. □

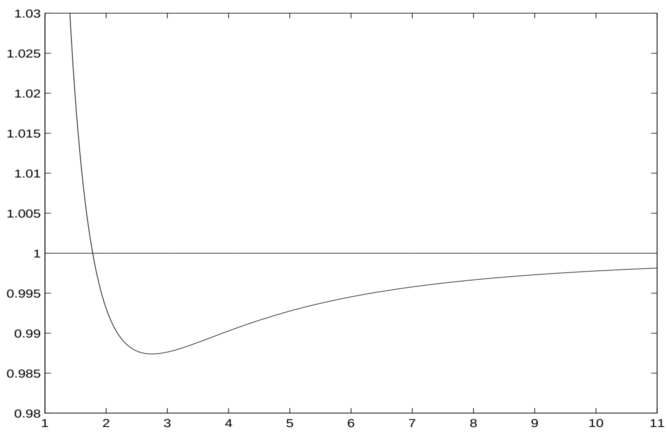

Figure 7: Plot of function $H(u)$, defined in (23). Important to note is that this function is smaller than 1 for $u > 1.7815$.

Corollary 1 *For $\lambda \to \infty$,*

$$\frac{\partial r}{\partial \lambda}(\frac{\sigma^2}{2\lambda}, \lambda)$$

is negative and tends to 0*, but not faster than*

$$-2\sigma^4 \frac{\phi(\lambda)}{\lambda^2}$$

Proof: This follows from the asymptotic behavior of $\frac{\partial r}{\partial \lambda}(0, \lambda)$ in Lemma 4, the fact that

$$\frac{\partial r}{\partial \lambda}(\frac{\sigma^2}{\lambda}, \lambda) \leq 0,$$

which follows from Theorem 2, and from the previous lemma, stating that $\frac{\partial r}{\partial \lambda}$ is convex between 0 and σ^2/λ. □

5.2 Intermediate results for the risk in one coefficient

This leaves us with the question to prove the inequality in (19). This section is purely technical, and may be skipped for understanding the rest of this paper.

Call

$$H(u) = u \ \frac{2 - \Phi_1(u + 1/u) - \Phi_1(u - 1/u)}{\phi_1(u + 1/u) + \phi_1(u - 1/u)}. \tag{23}$$

The plot of this function in Figure 7 seems to confirm that indeed $H(u) < 1$ for $u > 1.7815$. To make sure that this remains true for higher values of u, we start with the following lemma:

Lemma 8 *The function*

$$H_0(x) = \frac{1-\Phi_1(x)}{\phi_1(x)} \tag{24}$$

tends to zero for $x \to \infty$ and decreases monotonically for all positive x.

Proof:
The computation of $\lim_{x\to\infty} H_0(x)$ is straightforward, using De L'Hôpital's rule.

Next, it can be verified that $H_0(x)$ satisfies the first order differential equation:

$$H_0'(x) = xH_0(x) - 1,$$

and so

$$H_0''(x) = xH_0'(x) + H_0(x).$$

Suppose $H_0'(x) > 0$. Since $H_0(x) > 0$, this means that $H_0''(x) > 0$. So, $H_0'(x)$ would be positive and increasing. This conflicts with the limit of $H(x)$ to be zero. □

As a consequence of this lemma, and since it follows from $a/b \le c/d$ that $a/b \le (a+c)/(b+d) \le c/d$, we have that for positive u:

$$u\,\frac{1-\Phi_1(u+\frac{1}{u})}{\phi_1(u+\frac{1}{u})} \le u\,\frac{\left(1-\Phi_1(u+\frac{1}{u})\right)+\left(1-\Phi_1(u-\frac{1}{u})\right)}{\phi_1(u+\frac{1}{u})+\phi_1(u-\frac{1}{u})} \le u\,\frac{1-\Phi_1(u-\frac{1}{u})}{\phi_1(u-\frac{1}{u})}.$$

It is easy to verify that both the left and the right of this inequalities tend to one if $u \to \infty$, and so, we may conclude that

$$\lim_{u\to\infty} H(u) = 1. \tag{25}$$

We now use the fact that

$$\phi_1(u-\frac{1}{u}) = \mathrm{e}^2\,\phi_1(u+\frac{1}{u})$$

to rewrite $H(u)$ as:

$$H(u) = \frac{u}{\mathrm{e}^2+1}\,\frac{2-\Phi_1(u+\frac{1}{u})-\Phi_1(u-\frac{1}{u})}{\phi_1(u+\frac{1}{u})}.$$

From this expression we calculate:

$$H'(u) = -u - \frac{\mathrm{e}^2-1}{\mathrm{e}^2+1}\frac{1}{u} + \left(u+\frac{1}{u}-\frac{1}{u^3}\right)H(u).$$

This allows us to prove that:

Lemma 9 *The function $H(u)$, as defined in Equation (23), satisfies:*

$$H(u) \leq 1, \quad \forall u \geq \sqrt{\frac{e^2+1}{2}} \approx 2.048. \tag{26}$$

Proof:
Suppose $H(u) \geq 1$, then

$$H'(u) \geq \frac{e^2+1}{2}\frac{1}{u} - \frac{1}{u^3}.$$

This expression is positive for $u \geq \sqrt{\frac{e^2+1}{2}}$, which means that $H(u)$ would increase and its limit could never become one. $\square$

5.3 Piecewise smooth functions

If the noise-free signal is an exact polynomial, and if the multiresolution analysis has sufficiently many vanishing moments, the signal can be written as a linear combination of scaling functions at an arbitrary resolution. This means that all detail coefficients, i.e., the wavelet coefficients, are exactly zero. We used this property to describe what happens with *piecewise* polynomials. To investigate piecewise smooth functions, we follow the same way: we start with properties for wavelet coefficients of functions with a certain degree of smoothness. Of course, these coefficients will not be exactly zero, but smooth functions can be approximated by polynomials and this approximation guarantees that wavelet coefficients are "sufficiently" small. All this motivates the following definition of *Lipschitz continuity* as a measure of smoothness:

Definition 2 *A function f is called (uniformly) Lipschitz α over an interval $[a, b]$ if for all $x \in [a, b]$ there exists a polynomial $p_x(t)$, and there exists a constant K, independent of x, so that*

$$\forall t \in [a, b] : |f(t) - p_x(t)| \leq K|t - x|^{\alpha}. \tag{27}$$

A Lipschitz continuous function can be locally approximated by a polynomial. The effect on the wavelet coefficients of such a function is described by the following theorem, due to Jaffard [8, 12]:

Theorem 3 *If a function f is uniformly Lipschitz α over $[a, b]$ and if the wavelet function ψ has p vanishing moments with $p \geq \alpha$, then*

$$\exists C \in \mathbb{R}^+, \forall j \in \mathbb{Z}, \forall i = 1 \ldots 2^j : |\langle f, \psi_{j,k} \rangle| \leq C 2^{-j\left(\alpha + \frac{1}{2}\right)}. \tag{28}$$

We now have all the elements to study the asymptotic behavior of the minimum MSE threshold λ^* for piecewise smooth signals, corrupted by white, stationary and Gaussian noise. We work on a bounded interval and assume that the number of singularities is finite. Like in the piecewise polynomial case, we call I_0 the set of coefficients corresponding to basis functions not interfering with one the singularities and I_1 all the other coefficients. The cardinal numbers of these sets are respectively M_0 and M_1.

We call v^* the critical untouched coefficient value, corresponding to the minimum MSE threshold λ^*: for noise-free coefficients below this value, λ^* is too small, for coefficients with larger magnitude, the threshold is too large. The minimum MSE threshold is the best compromise between these two groups and in Section 5.1, Equation 20 we found a lower bound for the critical coefficient value: $v^* \geq \sigma^2/\lambda^* \geq \sigma^2/2\lambda^*$. This means that if we call

$$F_0 = \{i = 1, \ldots, N \mid |v_i| \leq v^*\},$$

and

$$F_L = \{i = 1, \ldots, N : |v_i| \leq \sigma^2/2\lambda^*\},$$

then we know that

$$F_L \subset F_0.$$

We call $K_0 = \#F_0$ the number of coefficients beneath the critical value and $K_1 = N - K_0$ the number of coefficients above this value. It is important to note that

$$v_i \in F_0 \Leftrightarrow \frac{\partial r}{\partial \lambda}(v_i, \lambda^*) \leq 0,$$

and so, we can write the equation

$$\mathrm{E}R'(\lambda^*) = 0$$

as

$$-\sum_{i \in F_0} \frac{\partial r}{\partial \lambda}(v_i, \lambda^*) = \sum_{i \in F_1} \frac{\partial r}{\partial \lambda}(v_i, \lambda^*), \tag{29}$$

and both sides in this equation have only positive terms.

We suppose that the coefficients are computed from a direct projection of the continuous signal:

$$v_i = \sqrt{N}\langle f, \psi_{j,k}\rangle = 2^{J/2}\langle f, \psi_{j,k}\rangle.$$

In practice, these values are approximated by a Fast Wavelet Transform on sample values, or pre-filtered sample values.

To have an idea of the asymptotic behavior of the sums in Equation (29), we count the number of terms on the left-hand side:

$$K_0 = \#F_0 \geq \#(F_0 \cap I_0) \geq \#(F_L \cap I_0).$$

The coefficients at the jth resolution level in I_0 satisfy $v_i \leq C 2^{J/2} 2^{-j(\alpha+1/2)}$. So, if a given resolution level j satisfies

$$C 2^{(J-2j\alpha-j)/2} \leq \sigma^2/2\lambda^*,$$

we are sure that all I_0-coefficients at that level are in $F_L \subset F_0$. This condition on j can be worked out as:

$$j \geq \frac{J - \frac{2}{\log 2}\log \sigma^2 + \frac{2}{\log 2}\log \lambda^* + \frac{2}{\log 2}\log 2C}{2\alpha+1}.$$

In the expression on the right-hand side λ^* is expected to depend on J, but we assume that $\log \lambda^*$ can be neglected with respect to J. If this is not the case, this means that the optimal threshold would increase at least linearly with N. This would not pose any problem to our further analysis, but it is rather unlikely to happen, apart from some pathological cases (a zero signal, for instance). We also drop the constant terms in this right-hand side and we express that $K_1 = N - K_0$ must be smaller than the total number of coefficients at scales j not satisfying this condition *plus* the total number of coefficients in I_1:

$$\begin{aligned} K_1 &\leq M_1 + \sum_{j=0}^{\lfloor J/(2\alpha+1)\rfloor} 2^j \\ &\approx M_1 + \frac{2^{J/(2\alpha+1)+1} - 1}{2-1} \\ &\sim 2 \cdot 2^{J/(2\alpha+1)} \\ &= 2N^{1/(2\alpha+1)}. \end{aligned}$$

So

$$\frac{K_1}{N} \sim 2N^{1/(2\alpha+1)-1} = 2N^{-2\alpha/(2\alpha+1)}.$$

Taking the logarithm of these asymptotics gives:

$$\begin{aligned} \log K_1 &\sim \frac{\log N}{2\alpha+1}, \\ \log K_0 &= \log(N - K_1) = \log N + \log\left(1 - \frac{K_1}{N}\right) \sim \log N - \frac{K_1}{N} \sim \log N. \end{aligned}$$

Actually, we have started from a lower bound for K_0 to find this behavior. Obviously $\log K_0$ cannot grow faster than $\log N$, since $K_0 \leq N$. On the other hand, the behavior of $\log K_1$ is based on an upper bound. Theoretically, $\log K_1$ may grow slower than $\log N/(2\alpha+1)$. Following the analysis below, it would turn out that in that case, the minimum MSE threshold would grow (a little) faster. The asymptotic behavior that we will find is a minimal one.

We are now ready to fill in both sides of Equation (29). For the right-hand side, we assume that λ^* increases slower than v_i, and from Lemma 3 it then follows that this side behaves like $K_1 2\lambda^*$. For the left-hand side, we use lower bounds, both for the number of coefficients in F_0 as for their asymptotic behavior. We only consider the coefficients in $F_L \subset F_0$ for which Corollary 1 gives a lower bound on the asymptotic behavior.

$$
\begin{aligned}
K_0 2\sigma^4 \frac{\phi(\lambda)}{\lambda^2} &\sim K_1 2\lambda \\
\sigma^4 \frac{\phi(\lambda)}{\lambda^3} &\sim \frac{K_1}{N - K_1} \\
\sigma^4 \frac{\phi(\lambda)}{\lambda^3} &\sim \frac{1}{\frac{N}{K_1} - 1} \\
\sigma^4 \frac{\phi(\lambda)}{\lambda^3} &\sim \frac{1}{\frac{1}{2} N^{2\alpha/(2\alpha+1)} - 1} \\
\sigma^4 \frac{\mathrm{e}^{-\lambda^2/2\sigma^2}}{\sqrt{2\pi}\sigma\lambda^3} &\sim 2N^{-2\alpha/(2\alpha+1)}.
\end{aligned}
$$

Taking the logarithm on both sides leads to the following theorem:

Theorem 4 *If a function f is Lipschitz α on $[0,1]$, except in a finite number of points and the wavelet analysis has p vanishing moments with $p \geq \alpha$, then the minimum MSE-threshold λ^* for denoising the corrupted observation*

$$y_i = f(i/N) + \eta_i \qquad i = 1, \ldots, N$$

behaves asymptotically as

$$\lambda^* \sim \sqrt{\frac{2\alpha}{2\alpha+1}} \sqrt{2 \log N}\sigma, \tag{30}$$

if the number of observations N increases.

The factor

$$\sqrt{\frac{2\alpha}{2\alpha+1}}$$

comes from the fact that for piecewise smooth functions, coefficients with no interaction with the singularities are not exactly zero. In our analysis, when estimating the number of coefficients K_1 above the critical value v^*, we even neglected the singularity coefficients, compared to these non-zero coefficients with no singularity interaction: the same behavior would appear (as a lower bound) for signals with no singularity at all.

6 Conclusion

We have proven that the minimum risk threshold is slowly growing if the sample size increases. For piecewise polynomials, the minimum risk threshold asymptotically coincides with the universal threshold, for general piecewise smoothness, the minimum risk threshold is lower, but it comes close to the universal threshold within a constant factor.

Acknowledgement

This paper presents research results of the Belgian Programme on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture. The scientific responsibility rests with its authors. The first author is financed by a grant from the Flemish Institute for the Promotion of Scientific and Technological Research in the Industry (IWT).

References

[1] A. Chambolle, R. A. DeVore, N.-Y. Lee, and B. J. Lucier. Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage. *IEEE Transactions on Image Processing*, 7(3):319–355, March 1998.

[2] D. L. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, May 1995.

[3] D. L. Donoho and I. M. Johnstone. Neo-classical minimax theorems, thresholding, and adaptation. Preprint, presented at the Fifth Symposium on Statistical Decision Theory, June 1992.

[4] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[5] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90:1200–1224, 1995.

[6] P. Hall and P. Patil. Formulae for mean integrated squared error of non-linear wavelet-based density estimators. *Annals of Statistics*, 23(3):905–928, 1995.

[7] P. Hall and P. Patil. On the choice of smoothing parameter, threshold and truncation in nonparametric regression by non-linear wavelet methods. *Journal of the Royal Statistical Society, Series B*, 58(2):361–377, 1996.

[8] S. Jaffard. Pointwise smoothness, two-microlocalisation and wavelet coefficients. *Publicacions Matemàtiques*, 35:155–168, 1991.

[9] M. Jansen and A. Bultheel. Multiple wavelet threshold estimation by generalized cross validation for images with correlated noise. *IEEE Transactions on Image Processing*, 8(7):947–953, July 1999.

[10] M. Jansen, M. Malfait, and A. Bultheel. Generalized cross validation for wavelet thresholding. *Signal Processing*, 56(1):33–44, January 1997.

[11] I. M. Johnstone and B. W. Silverman. Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Series B*, 59:319–351, 1997.

[12] S. G. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 525 B Street, Suite 1900, San Diego, CA, 92101-4495, USA, 1998.

[13] G. P. Nason. Wavelet shrinkage using cross validation. *Journal of the Royal Statistical Society, Series B*, 58:463–479, 1996.