Parametric modelling of thresholds across scales in wavelet regression

Anestis Antoniadis[*] Piotr Fryzlewicz[†]

May 9, 2005

Abstract

We propose a parametric wavelet thresholding procedure for estimation in the "function plus independent, identically distributed Gaussian noise" model. To reflect the decreasing sparsity of wavelet coefficients from finer to coarser scales, our thresholds also decrease. They retain the noise-free reconstruction property while being lower than the universal threshold, and are jointly parameterized by a single scalar parameter. We show that our estimator achieves near-optimal risk rates for the usual range of Besov spaces. We propose a cross-validation technique for choosing the parameter of our procedure. A simulation study demonstrates a very good performance of our estimator compared to other state-of-the-art techniques. We discuss an extension to non-Gaussian noise.

Keywords: Asymptotic rates, Besov spaces, noise-free reconstruction, thresholding, wavelet decomposition.

1 Introduction

We are studying the classical nonparametric regression problem of recovering the values of an unknown function $f : [0,1] \mapsto \mathbb{R}$ from noisy observations on an equidistant grid:

$$y_i = f(i/n) + \varepsilon_i, \quad i = 1, \ldots, n = 2^J, \tag{1}$$

where ε_i are independent and distributed as $N(0, \sigma^2)$. We are concerned with estimators $\hat{f}$ based on wavelets: for an overview of wavelet methods in statistics, see e.g. Vidakovic (1999). Given an orthonormal Discrete Wavelet Transform (DWT) $W : \mathbb{R}^n \mapsto \mathbb{R}^n$, denote

[*] Laboratoire de Modelisation et Calcul, Université Joseph Fourier, Tour IRMA, B.P.53, 38041 Grenoble CEDEX 9, France.

[†] Author for correspondence. Until 31 August 2005: Department of Mathematics, South Kensington Campus, Imperial College London, London SW7 2AZ, UK; email: p.fryzlewicz@imperial.ac.uk. From 1 September 2005: Department of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW, UK.

the vector of noisy wavelet coefficients by $y_{j,k} = Wy$, the vector of noise-free wavelet coefficients by $d_{j,k} = Wf$, the vector of estimated wavelet coefficients by $\hat{d}_{j,k} = W\hat{f}$, and the vector of "wavelet noise" coefficients by $\varepsilon_{j,k} = W\varepsilon$, where $j = 0$ ($j = J - 1$) is the coarsest (finest) detail scale. At any given scale j, the detail coefficients are indexed by $k = 1, \ldots, 2^j$. The only smooth coefficient is indexed by $(j, k) = (-1, 1)$. By the linearity of the wavelet transform W, in the wavelet domain (1) becomes

$$y_{j,k} = d_{j,k} + \varepsilon_{j,k}, \tag{2}$$

where, due to the orthonormality of W, the $\varepsilon_{j,k}$'s are again independent $N(0, \sigma^2)$. In function estimation via wavelets, each $y_{j,k}$ is used to obtain an estimate $\hat{d}_{j,k}$ of $d_{j,k}$. The estimate $\hat{f}$ is then obtained upon applying the inverse wavelet transform W^{-1} to $\hat{d}_{j,k}$.

For many signals, the representation (2) is sparse, i.e. only a few true coefficients $d_{j,k}$ are significantly different from zero. Motivated by this, Donoho and Johnstone (1994) proposed thresholding as a way of estimating $d_{j,k}$ from $y_{j,k}$. Thresholding annihilates those empirical coefficients $y_{j,k}$ which fall below a certain threshold t, and, provided that t is chosen "correctly", is an extremely effective denoising technique despite its simplicity.

For the universal threshold $t = \sigma(2 \log n)^{1/2}$ (Donoho and Johnstone (1994)), the following noise-free reconstruction property holds: when the true signal f is constant, then, with high probability, the estimate $\hat{f}$ is also constant and equal to the empirical mean of $\{y_i\}_{i=1}^n$. This is a desirable property, for example, in wavelet-based functional analysis of variance tests (Abramovich et al. (2004)). It ensures that the reconstruction is visually appealing as, asymptotically, it contains no noise. The universal threshold is asymptotically the lowest scale-independent threshold which satisfies the noise-free reconstruction property, but it is still "too high" in the sense that its application often leads to oversmoothing.

Motivated by the often observed decreasing sparsity of wavelet coefficients from finer to coarser scales, some authors have proposed scale-dependent thresholds t_j which often achieve better mean-square performance than scale-independent thresholds t, see e.g. Johnstone and Silverman (2005). In this paper, we combine these two important issues: the noise-free reconstruction property, and scale-dependent thresholding. We investigate whether it is possible to devise a scale-dependent thresholding scheme t_j which performs better than the universal threshold in the mean-square sense, but still retains the noise-free reconstruction property. Moreover, we are particularly interested in the case where we can impose some parametric dependence between the threshold values $\{t_j\}_{j=0}^{J-1}$, i.e. assume $t_j = t_\theta(j)$, where t_θ is a family of functions parameterized by a single scalar parameter θ. The rationale here is that by choosing t_j "jointly", and not separately for each scale, we can potentially obtain a stable selection procedure even for coarser scales where only a few wavelet coefficients are available. The function t_θ will often be referred to as a "threshold profile".

The paper is organised as follows: in Section 2, we consider a general noise-free reconstruction property of scale-dependent thresholding, leading to a specific family of threshold profiles. In Section 3, we demonstrate the mean-square near-optimality of the new thresholding procedure over a range of Besov spaces. In Section 4, we introduce a data-driven technique for selecting the parameter θ of our procedure. In Section 5, the performance of the new method is investigated in a simulation study. In Section 6, we show how to extend the proposed methodology to non-Gaussian noise distributions.

2 A generic noise-free reconstruction property

The estimator considered in this paper is the hard thresholding estimator

$$\hat{d}_{j,k}(t_j) = y_{j,k}\,\mathbb{I}\{|y_{j,k}| > t_j\}, \qquad (3)$$

for $j = 0, \dots, J-1$ and $k = 1, \dots, 2^j$. However, results analogous to those obtained in this paper can also be derived for the soft thresholding case. We skip this case for simplicity, and due to the inferior practical performance of soft thresholding estimators, see e.g. Antoniadis et al. (2001). We leave the smooth coefficient unchanged: $\hat{d}_{-1,1} = y_{-1,1}$. For notational simplicity, we assume $\sigma = 1$ throughout the paper. In practice, the parameter σ is often estimated from the data via the Median Absolute Deviation (MAD) estimator on the finest resolution level $J-1$.

The normality, independence and identical distribution of the "wavelet noise" coefficients $\varepsilon_{j,k}$ are the key ingredients of the denoising-via-thresholding theory due to Donoho and Johnstone (1994). In particular, the popular universal thresholding procedure is based on the following relation for independent, identically distributed standard normal variables $\varepsilon_{j,k}$:

$$\mathrm{pr}\left\{\max_{j=0,\dots,J-1;k=1,\dots,2^j} |\varepsilon_{j,k}| > (2\log n)^{1/2}\right\} \to 0 \quad \text{as } n \to \infty. \qquad (4)$$

Using (4), it can be shown that applying the scale-independent threshold $t_j = t = (2\log n)^{1/2}$ in (3) leads to the noise-free reconstruction property: if f is a constant signal, then, with high probability, $\hat{f}$ is also constant and equal to the empirical mean of $\{y_i\}_{i=1}^n$. It can also be demonstrated that the choice $t = (2\log n)^{1/2}$ yields near-optimal Mean-Square Error (MSE) rates over a range of signal smoothness classes, and produces visually appealing reconstructions even for relatively small sample sizes n. However, it is well known that the universal threshold oversmooths: for non-zero signals f, too much signal gets killed in the process of thresholding. There arises a need for lower thresholds; however, replacing $t_j = t = (2\log n)^{1/2}$ in (3) with $t_j = t = (a\log n)^{1/2}$ for $a < 2$ ruins the noise-free reconstruction property as

$$\mathrm{pr}\left\{\max_{j=0,\dots,J-1;k=1,\dots,2^j} |\varepsilon_{j,k}| > (a\log n)^{1/2}\right\} \not\to 0 \quad \text{as } n \to \infty$$

if $a < 2$. Thus, the only way of obtaining thresholds which are lower than the universal threshold $t = (2\log n)^{1/2}$, but, possibly, still preserve the noise-free reconstruction property, is to resort to scale-dependent thresholds t_j.

As mentioned in Section 1, one other motivation for using scale-dependent thresholds t_j is the fact that for many signals, the coarser the scale, the larger the proportion of $d_{j,k}$'s which significantly differ from zero. By estimating those $d_{j,k}$'s as zero, we would unnecessarily kill significant information, and to prevent this, the use of lower threshold should be considered at coarses scales. Indeed, the fact that the sparsity often decreases from finer to coarser scales suggests using thresholding profiles which also decrease.

In what follows, we derive a sufficient condition for scale-dependent thresholds which (a) are lower than the universal threshold and decrease from finer to coarser scales, and (b) preserve the noise-free reconstruction property. Let $y_{j,k}$ denote wavelet coefficients of a "pure Gaussian noise" signal and assume that possibly different thresholds t_j are applied at each scale $j = 0, \ldots, J-1$. It can easily be shown that the noise-free reconstruction property occurs if and only if

$$\mathrm{pr}\left(|y_{J-1,1}| > t_{J-1} \vee \ldots \vee |y_{J-1,2^{J-1}}| > t_{J-1} \vee \ldots \vee |y_{0,1}| > t_0\right) \to 0 \quad \text{as} \quad n \to \infty.$$

Denoting the pdf (cdf) of a standard normal by ϕ (Φ), we obtain

$$\mathrm{pr}\left(|y_{J-1,1}| > t_{J-1} \vee \ldots \vee |y_{J-1,2^{J-1}}| > t_{J-1} \vee \ldots \vee |y_{0,1}| > t_0\right) \leq$$
$$\sum_{j,k} \mathrm{pr}\left(|y_{j,k}| > t_j\right) = \sum_{j=0}^{J-1} 2^j 2\{1 - \Phi(t_j)\} \leq 2 \sum_{j=0}^{J-1} 2^j \phi(t_j)/t_j.$$

Thus, a sufficient condition for the "noise-free reconstruction" property is

$$\lim_{J \to \infty} \sum_{j=0}^{J-1} \phi(t_j) 2^j / t_j = 0. \tag{5}$$

We assume that our thresholds are of the form $t_j = (2 \log n)^{1/2} t_\theta(j/(J-1))$, where $t_\theta(z) : [0,1] \mapsto [\delta, 1]$ is a family of continuous, nondecreasing functions with $\delta > 0$. Note that setting $t_\theta(z) \equiv 1$ yields the classical universal threshold. Continuing from (5), we have

$$\sum_{j=0}^{J-1} \frac{\phi(t_j) 2^j}{t_j} = \frac{1}{(4\pi J \log 2)^{1/2}} \sum_{j=0}^{J-1} \frac{2^{-J t_\theta^2 \left(\frac{j}{J-1}\right) + j}}{t_\theta\left(\frac{j}{J-1}\right)} \leq \frac{1}{\delta (4\pi J \log 2)^{1/2}} \sum_{j=0}^{J-1} 2^{-J t_\theta^2 \left(\frac{j}{J-1}\right) + \frac{Jj}{J-1}}.$$

As $J \to \infty$, it suffices to investigate when the sum is bounded in J. The sum behaves like

$$J \int_0^1 2^{J\{x - t_\theta^2(x)\}} dx. \tag{6}$$

If $t_\theta^2(x) \leq x$ on any set of non-zero measure in $[0,1]$, then (6) is not bounded. Of course, we cannot speak here of the "smallest permitted" $t_\theta^2(x)$, as any such that $t_\theta^2(x) \geq \delta$ and $t_\theta^2(x) > x$ a.e. will do, but for simplicity we single out "almost the smallest permitted" $t_\theta^2(x)$ of the form $t_\delta^2(x) = \delta + (1-\delta)x$, which is a natural lower boundary for the family of functions $t_\theta(x) = \{\theta + (1-\theta)x\}^{1/2}$, parameterized by a one-dimensional parameter $\theta \in [\delta, 1]$. The expression (6) is bounded for any $t_\theta(x)$ of this specific form and thus the threshold profile

$$t_j = (2 \log n)^{1/2} \left\{\theta + (1-\theta) \frac{j}{J-1}\right\}^{1/2} \tag{7}$$

preserves the noise-free reconstruction property for any $\theta \in [\delta, 1]$. Motivated by this result, we propose to estimate $d_{j,k}$ by the hard thresholding estimator (3) with t_j as in (7). Due to the particular form of this threshold profile, we label the new estimator "SQRT".

3 Risk properties of the SQRT estimator

In this section, we consider the MSE properties of the SQRT estimator. We assume that the unknown signal f belongs to a Besov ball of radius $C > 0$ on $[0, 1]$, $B^{\nu}_{p,q}(C)$, where $\nu > 0$ and $0 < p, q \leq \infty$. Roughly speaking, the not necessarily integer parameter ν indicates the number of derivatives of f, where their existence is required in the L^p-sense, and thus p can be viewed as the measure of inhomogeneity of f. The additional parameter q provides a further finer gradation. Besov classes include the traditional Hölder and Sobolev classes of smooth functions ($p = q = \infty$ and $p = q = 2$, respectively) and various classes of spatially inhomogeneous functions like the class of functions of bounded variation, sandwiched between $B^1_{1,\infty}$ and $B^1_{1,1}$. Also note that if the father and mother wavelets have regularity $r > 0$, then the corresponding wavelet basis is an unconditional basis for the Besov spaces $B^{\nu}_{p,q}([0, 1])$ for $0 < r\nu < r$, $0 < p, q \leq \infty$. This allows one to characterise Besov balls in terms of the wavelet coefficients $d'_{j,k} = d_{j,k}/n^{1/2}$ of the function f in the following way. Define the Besov sequence ball of radius C as

$$b^{\nu}_{p,q}(C) = \left\{ d'_{j,k} : \sum_{j=0}^{\infty} 2^{jsq} \|d'_j\|^q_p \leq C^q \right\},$$

where $s = \nu + 1/2 - 1/p$ and $\|d'_j\|^p_p = \sum_{k=1}^{2^j} |d'_{j,k}|^p$. The membership of f in $B^{\nu}_{p,q}(C)$ can be thought of as being equivalent to the membership of $\{d'_{j,k}\}_{j,k}$ in $b^{\nu}_{p,q}(C)$. The reader is referred to Meyer (1992) for rigorous definitions and a detailed study of Besov spaces.

The following theorem establishes the MSE near-optimality of our SQRT estimator over a wide range of Besov sequence spaces.

Theorem 3.1 *Given the regression problem (1), let $\hat{f}$ be the SQRT estimator of f, constructed by applying the inverse DWT to the sequence of estimated wavelet coefficients $\hat{d}_{j,k}(t_j)$ with thresholds t_j defined by (7), for any fixed $\theta \in [\delta, 1]$. Denote $\hat{d}'_{j,k}(t) = \hat{d}_{j,k}(t)/n^{1/2}$. If $0 < p, q \leq \infty$ and $\nu > 1/p$, then*

$$\begin{aligned} \sup_{d'_{j,k} \in b^{\nu}_{p,q}(C)} \mathrm{MSE}(\hat{f}, f) &= \sup_{d'_{j,k} \in b^{\nu}_{p,q}(C)} \frac{1}{n} \sum_{i=1}^{n} E\left\{ \hat{f}(i/n) - f(i/n) \right\}^2 \\ &= \frac{\sigma^2}{n} + \sup_{d'_{j,k} \in b^{\nu}_{p,q}(C)} \sum_{j=0}^{J-1} \sum_{k=1}^{2^j} E\left\{ \hat{d}'_{j,k}(t_j) - d'_{j,k} \right\}^2 \leq C_0 n^{-\frac{2\nu}{2\nu+1}} \log n, \end{aligned}$$

where C_0 is independent of n.

The rate $O(n^{-\frac{2\nu}{2\nu+1}})$ is the best possible MSE rate for Besov spaces, and SQRT achieves it up to the logarithmic term: hence the name "near-optimality". The above rate is identical to that achieved by the classical universal thresholding estimator. The proof of Theorem 3.1 appears in the Appendix.

4 Data-driven choice of θ

Even though the universal threshold, being a special case of SQRT with $\theta = 1$, is known to oversmooth, it is not true that "for any signal, the lower the value of θ, the better the reconstruction". As a counterexample, consider the zero signal, where the MSE of the SQRT decreases with θ increasing: clearly, for the zero signal, the higher the thresholds, the better the reconstruction. Thus, there arises a need for a data-driven choice of θ.

In this section, we describe a computational procedure, based on the leave-half-out cross-validation method of Nason (1996), for choosing a suitable value of θ from the data. Given the value θ_l from a pre-selected grid $\{\theta_l\}_{l=1}^{L}$, we split the data $\{y_i\}_{i=1}^{n}$ into the odd subsample $\{y_{2i-1}\}_{i=1}^{n/2}$ and the even subsample $\{y_{2i}\}_{i=1}^{n/2}$. We then run the SQRT algorithm with parameter θ_l on the two subsamples to obtain the odd and even estimates, respectively. Finally, we measure the distance between the odd estimate and the even subsample, and add it to the distance between the even estimate and the odd subsample. The selected value of θ_l is the one which minimises the sum of these two distances.

The version of our SQRT algorithm which includes the above cross-validatory procedure for choosing θ is labelled SQRT-CV. The SQRT-CV algorithm is of computational order $O(Ln)$ and is fully automatic, i.e. no parameters need to be supplied by the user.

5 Empirical performance of SQRT and SQRT-CV

In this section, we compare the finite-sample performance of the SQRT and SQRT-CV estimators to a selection of other wavelet denoising methods. Our test functions are Donoho and Johnstone's bumps, doppler, heavisine and blocks, as well as the zero signal, sampled at 1024 equispaced points. The standard deviation of the noise is always 1 but it is unknown to the estimation procedures and always estimated using MAD on the finest detail level. The respective root signal-to-noise ratios are: 1.33, 1.45, 2.97, 1.91, 0. Note that these signal-to-noise ratios are relatively low, i.e. the observed signals have a considerably noisy appearance. The respective analysing wavelets are: Daubechies' Extremal Phase (DEP) 2, Daubechies' Least Asymmetric (DLA) 9, DLA 8, DEP 1, and DLA 4. Periodic boundary conditions are assumed. Due to their superior performance, we only compare the Translation-Invariant (TI; Nason and Silverman (1995)) versions of the estimators. The competitors are:

UNIV-TI: TI universal hard thresholding with all levels thresholded. In a recent study assessing the empirical performance of various wavelet-based denosing methods (Antoniadis et al. (2001)), UNIV-TI consistently performed the best, or nearly the best, among various modern wavelet smoothing techniques.

EB-TI: the TI version of the empirical Bayes (eBayes) procedure of Johnstone and Silverman (2005). In the simulation study reported therein, eBayes is shown to outperform several state-of-the-art denoising techniques.

	UNIV-TI	EB-TI	SQRT-TI	SQRT-TI-CV
bumps	154	134	$\boxed{\boxed{126}}$	$\boxed{127}$
doppler	62	71	$\boxed{\boxed{57}}$	$\boxed{58}$
heavisine	45	$\boxed{41}$	$\boxed{\boxed{37}}$	$\boxed{41}$
blocks	80	87	$\boxed{\boxed{72}}$	$\boxed{\boxed{72}}$
zero	$\boxed{\boxed{24}}$	78	72	$\boxed{27}$

Table 1: ISE averaged over 100 sample paths (×1000 (except zero: ×10000) and rounded) for the 4 competing methods. Double box indicates best, and box — 2nd best result. See the discussion in Section 5.

SQRT-TI: the TI version of our SQRT estimator with hard thresholding and the profile defined by $t_{0.01}(x) = (0.01 + 0.99x)^{1/2}$.

SQRT-CV-TI: the TI version of our SQRT estimator with hard thresholding where θ is selected using the cross-validatory procedure of Section 4, over the grid $\theta_l = l/10$ for $l = 2, \ldots, 10$ and $\theta_1 = 0.01$.

The ISE for each method, averaged over 100 sample paths, is shown in Table 1. SQRT-CV-TI is clearly the preferred option here: except the zero signal where it is, naturally enough, slightly outperformed by UNIV-TI, it outperforms EB-TI by 0–18%, and UNIV-TI by 6–18%. Given the quality of the competitors, this is indeed a significant improvement.

The computational complexity of the SQRT-TI-CV algorithm is $O(n(L + \log n))$, where L is the size of the grid $\{\theta_l\}_{l=1}^{L}$. In practice, the software is fast, which is partly due to the fact that the threshold choice is straightforward and requires no computationally intensive procedures. The SQRT(-TI)(-CV) algorithm is easy to code in any package which implements the DWT.

6 Other noise distributions

In this section, we demonstrate how the proposed estimation method can be extended to non-Gaussian noise distributions. Our setup is

$$\tilde{y}_i = f(i/n) + \tilde{\varepsilon}_i, \quad i = 1, \ldots, 2^J, \tag{8}$$

where $\tilde{\varepsilon}_i$ are independent and identically distributed and follow a known distribution, not necessarily Gaussian, with $E(\tilde{\varepsilon}_i) = 0$. This is an additive setup where the noise $\tilde{\varepsilon}_i$ does not depend on the underlying signal f. In the wavelet domain, (8) becomes $\tilde{y}_{j,k} = d_{j,k} + \tilde{\varepsilon}_{j,k}$, where the notation is analogous to the Gaussian case. For a fixed j, each component of the vector $\{\tilde{\varepsilon}_{j,k}\}_{k=1}^{2^j}$ is identically distributed, and its distribution can be either derived analytically, or easily approximated numerically via Monte Carlo simulations, by performing the DWT of simulated vectors $\{\tilde{\varepsilon}_i\}_{i=1}^{n}$. Thus, in the remaining part of this section we assume that the distribution of $\tilde{\varepsilon}_{j,k}$ is known for each j. The noise-free reconstruction property arises if and only if

$$\mathrm{pr}(|\tilde{\varepsilon}_{J-1,1}| > \tilde{t}_{J-1} \vee \ldots \vee |\tilde{\varepsilon}_{J-1,2^{J-1}}| > \tilde{t}_{J-1} \vee \ldots \vee |\tilde{\varepsilon}_{0,1}| > \tilde{t}_0) \to 0$$

as $n \to \infty$. But we have

$$\begin{aligned} &\mathrm{pr}(|\tilde{\varepsilon}_{J-1,1}| > \tilde{t}_{J-1} \vee \ldots \vee |\tilde{\varepsilon}_{J-1,2^{J-1}}| > \tilde{t}_{J-1} \vee \ldots \vee |\tilde{\varepsilon}_{0,1}| > \tilde{t}_0) \\ &\leq \sum_{j,k} \mathrm{pr}(|\tilde{\varepsilon}_{j,k}| > \tilde{t}_j) = \sum_{j=0}^{J-1} 2^j \mathrm{pr}(|\tilde{\varepsilon}_{j,k}| > \tilde{t}_j). \end{aligned}$$

To ensure that our estimator has similar visual properties as in the Gaussian case, i.e. that a similar small proportion of the noise survives the thresholding, a natural requirement is that the individual exceedance probabilities $\mathrm{pr}(|\tilde{\varepsilon}_{j,k}| > \tilde{t}_j)$ should be the same as in the Gaussian case. In other words, we find $\tilde{t}_j$ by numerically solving the equations

$$\mathrm{pr}(|\tilde{\varepsilon}_{j,k}| > \tilde{t}_j) = \mathrm{pr}(|\varepsilon_{j,k}| > t_j) = 2\{1 - \Phi(t_j)\}, \quad j = 0, \ldots, J-1,$$

where t_j are the SQRT thresholds of the form (7), suitable for Gaussian data. This indeed guarantees that the noise-free reconstruction property holds as we have

$$\sum_{j=0}^{J-1} 2^j \mathrm{pr}(|\tilde{\varepsilon}_{j,k}| > \tilde{t}_j) = 2 \sum_{j=0}^{J-1} 2^j \{1 - \Phi(t_j)\} \leq 2 \sum_{j=0}^{J-1} 2^j \phi(t_j)/t_j,$$

and the latter quantity converges to zero by formula (5).

Mean-square risk analysis of wavelet thresholding estimators for non-Gaussian data is typically not straightforward, and, for the estimator proposed in this section, is beyond the scope of this short communication. However, we note that it can be performed using techniques as in Neumann and von Sachs (1995).

Acknowledgements

We are grateful to the Editor and an anonymous Referee for useful comments and suggestions which have significantly improved this paper.

A Proof of Theorem 3.1

The second equality is due to the orthonormality of the DWT. For n large enough such that $n^\theta \geq 4$, we apply the "oracle inequality" from Theorem 7 of Donoho and Johnstone (1994) to obtain

$$\sum_{j=0}^{J-1} \sum_{k=1}^{2^j} \mathbb{E}\left\{ \hat{d}'_{j,k}(t_j) - d'_{j,k} \right\}^2 \leq$$

$$\begin{aligned}
&\sum_{j=0}^{J-1}\sum_{k=1}^{2^j}\left[2.4+\left\{\theta+(1-\theta)\frac{j}{J-1}\right\}2J\log\,2\right]\\
&\times\left[2^{-J\{\theta+(1-\theta)\frac{j}{J-1}+1\}}+\min\left\{(d'_{j,k})^2,n^{-1}\right\}\right]=\\
&\frac{1}{2^J}\sum_{j=0}^{J-1}2^j\left[2.4+\left\{\theta+(1-\theta)\frac{j}{J-1}\right\}2J\log\,2\right]2^{-J\{\theta+(1-\theta)\frac{j}{J-1}\}}\\
&+\sum_{j=0}^{J-1}\sum_{k=1}^{2^j}\left[2.4+\left\{\theta+(1-\theta)\frac{j}{J-1}\right\}2J\log\,2\right]\min\left\{(d'_{j,k})^2,n^{-1}\right\}=I+II.
\end{aligned}$$

Note that I is at most of order

$$\frac{2.4+2J\log\,2}{2^J}\int_0^J 2^{x-J\{\theta+(1-\theta)\frac{x}{J-1}\}}dx\leq\frac{2.4+2J\log\,2}{2^J}2^{-1}2^{\frac{\theta J-1}{J-1}}=O\left(\frac{\log\,n}{n}\right),$$

which, incidentally, is of the same order as the corresponding quantity for the universal threshold:

$$\frac{2.4+2J\log\,2}{2^J}\sum_{j=0}^{J-1}2^{j-J}=O\left(\frac{J}{2^J}\right)=O\left(\frac{\log\,n}{n}\right).$$

We now focus on II. Since $\theta\in[\delta,1]$ and $j\leq J-1$, note that II is less than the corresponding quantity for the classical universal threshold, which is

$$(2.4+\log\,n)\sum_{j=0}^{J-1}\sum_{k=1}^{2^j}\min\left\{n^{-1},(d'_{j,k})^2\right\}. \qquad (9)$$

Thus, instead of II, we shall consider (9). As $b^{\nu}_{p,q}(C)\subset b^{\nu}_{p,\infty}(C)$ for all q, we only have to consider the case $d'_{j,k}\in b^{\nu}_{p,\infty}(C)$, so we can assume $\|d'_j\|_p\leq C2^{-js}$ for all j, where C is a generic constant. The following argument was considered e.g. in Johnstone and Silverman (1997). We need to consider the cases $p\leq 2$ and $p>2$ separately. For $p\leq 2$, we first note the simple inequality

$$\begin{aligned}
\min\{|a|^2,|b|^2\} &= \min\{|a|^p,|b|^p\}\,\min\{|a|^{2-p},|b|^{2-p}\}\leq|a|^{2-p}\min\{|a|^p,|b|^p\}\\
&= \min\{|a|^2,|a|^{2-p}|b|^p\}.
\end{aligned}$$

Applying it with $a=n^{-1/2}$, $b=d'_{j,k}$, we bound the double sum in (9) as follows:

$$\begin{aligned}
\sum_{j=0}^{J-1}\sum_{k=1}^{2^j}\min\{n^{-1},(d'_{j,k})^2\} &\leq \sum_{j=0}^{J-1}\sum_{k=1}^{2^j}\min\{n^{-1},|d'_{j,k}|^p\,n^{p/2-1}\}\\
&\leq \sum_{j=0}^{J-1}\min\{2^jn^{-1},C^p2^{-jsp}n^{p/2-1}\}. \qquad (10)
\end{aligned}$$

Note that $2^j n^{-1} \le C^p 2^{-jsp} n^{p/2-1}$ if and only if $j \le J^* := (2\log_2 C + J)/(2\nu+1)$. Observe that asymptotically, we always have $J^* < J$. Assuming that J^* is an integer (it has no impact on the rates), we split (10) into two parts

$$\sum_{j=0}^{J*-1} 2^j n^{-1} + \sum_{j=J^*}^{J-1} C^p 2^{-jsp} n^{p/2-1}.$$

The first part is a partial sum of an increasing geometric series so, without going into details, it is bounded from above by a multiple of $n^{-1}2^{J^*} = O(n^{-2\nu/(2\nu+1)})$. The second part is a tail of decreasing geometric series so it is bounded from above by a multiple of $n^{p/2-1}2^{-J^*sp} = O(n^{-2\nu/(2\nu+1)})$. This proves the rate for $p \le 2$.

For $p > 2$, first note that the Hölder inequality gives $\|d'_j\|_2^2 \le 2^{j(1-2/p)}\|d'_j\|_p^2$. With this in mind, we bound the double sum in (9) as follows:

$$\begin{aligned}\sum_{j=0}^{J-1}\sum_{k=1}^{2^j} \min\{n^{-1}, (d'_{j,k})^2\} &\le \sum_{j=0}^{J-1} \min\{2^j n^{-1}, \|d'_j\|_2^2\} \\ &\le \sum_{j=0}^{J-1} \min\{2^j n^{-1}, C^2 2^{-2js} 2^{j(1-2/p)}\}. \qquad (11)\end{aligned}$$

As before, note that $2^j n^{-1} \le C^2 2^{-2js} 2^{j(1-2/p)}$ if and only if $j < J^*$. Again splitting the sum in (11) into two, we obtain

$$\sum_{j=0}^{J*-1} 2^j n^{-1} + \sum_{j=J^*}^{J-1} C^2 2^{-2js} 2^{j(1-2/p)}.$$

As we have already noted, the first part behaves like $O(n^{-2\nu/(2\nu+1)})$. The second part is a decreasing geometric series, so it is bounded from above by a multiple of $2^{J^*(-2s+1-2/p)} = O(n^{-2\nu/(2\nu+1)})$. This proves the desired rate for $p > 2$. □

References

Abramovich, F., Antoniadis, A., Sapatinas, T. & Vidakovic, B. (2004). Optimal testing in a fixed-effects functional analysis of variance model. *International Journal of Wavelets, Multiresolution and Information Processing*, **2**, 323–350.

Antoniadis, A., Bigot, J. & Sapatinas, T. (2001). Wavelet estimators in nonparametric regression: a comparative simulation study. *Journal of Statistical Software*, **6**, 1–83.

Donoho, D. L. & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

Johnstone, I. M. & Silverman, B. W. (1997) Wavelet threshold estimators for data with correlated noise. *J. Roy. Statist. Soc. Ser. B*, **59**, 319–351.

Johnstone, I. M. & Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *Annals of Statistics*, **33**, to appear.

Meyer, Y. (1992). *Wavelets and Operators.* Cambridge University Press.

Nason, G. P. (1996). Wavelet shrinkage using cross-validation. *J. Roy. Statist. Soc. Ser. B*, **58**, 463–479.

Nason, G. P. & Silverman, B. W. (1995). The stationary wavelet transform and some statistical applications. In A. Antoniadis and G. Oppenheim, editors, *Lecture Notes in Statistics, vol. 103*, pages 281–300. Springer-Verlag.

Neumann, M. & von Sachs, R. (1995). Wavelet thresholding: beyond the Gaussian i.i.d. situation. In A. Antoniadis and G. Oppenheim, editors, *Lecture Notes in Statistics, vol. 103*, pages 301–329. Springer-Verlag.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets.* Wiley.