DOI 10.1007/s10958-018-4048-x
Journal of Mathematical Sciences, Vol. 234, No. 6, November, 2018

THE ASYMPTOTIC BEHAVIOR OF THE OPTIMAL THRESHOLD MINIMIZING THE PROBABILITY-OF-ERROR CRITERION

A. A. Kudryavtsev[1] and O.V. Shestakov[1,2]

In this paper we consider the problem of estimation of a signal function from the noised observations via thresholding its wavelet coefficients. We find the asymptotic order of the optimal threshold that minimizes the probability of the maximum error between the estimates and the true wavelet coefficients exceeding a critical value.

1. Introduction

Estimation of a signal function using wavelet methods is a problem that has been drawing great attention over the last two decades. Many theoretical and practical results are introduced with an emphasis on problems of the choice of the parameters of these methods. The basic idea behind wavelet estimation is to get a relatively small number of wavelet coefficients to represent the underling signal function. A value called the threshold is used to remove or keep the wavelet coefficient. Hence, estimation quality depends on how efficient the threshold value is chosen.

There are a number of different strategies for choosing a threshold value [1–5]. Most of them are oriented at the minimization of the mean squared error (the risk) or its estimate. Statistical properties of this estimate were investigated in detail in [6] and [7]. In [8] the authors proposed a new cost function based on the probabilities of errors between the estimates and the true coefficients exceeding a critical value. They considered this cost function for each wavelet coefficient separately. In this paper we propose a generalization of the cost function from [8], which calculates the probability of the maximum error between the estimates and the true wavelet coefficients exceeding a critical value. We also investigate the asymptotic behavior of the minimax threshold value in the class of Lipschitz functions.

2. Statement of the problem

The wavelet decomposition of a signal function $f \in L^2(\mathbf{R})$ is the series

$$f = \sum_{j,k\in\mathbf{Z}} \langle f, \psi_{j,k}\rangle \psi_{j,k}, \tag{1}$$

where $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$ and ψ is a wavelet (the family $\{\psi_{j,k}\}_{j,k\in\mathbf{Z}}$ forms an orthonormal basis in $L^2(\mathbf{R})$). The index j in (1) is the scale, and the index k is the shift. The function ψ cannot be arbitrary, but it can possess some useful properties. For example, it can be M times differentiable and have M vanishing moments:

$$\int_{-\infty}^{\infty} x^k \psi(x)dx = 0, \quad k = 0, \ldots, M-1.$$

In what follows we consider the signal functions $f \in L^2(\mathbf{R})$ on a finite interval $[a, b]$, which are uniformly Lipschitz with some exponent $\gamma > 0$ and a Lipschitz constant $L > 0$: $f \in \mathrm{Lip}(\gamma, L)$. Suppose

[1] Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Moscow, Russia, e-mail: `oshestakov@cs.msu.su`
[2] Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia.

Proceedings of the XXXIII International Seminar on Stability Problems for Stochastic Models, Svetlogorsk, Russia, June 12–18, 2016

 1072-3374/18/2346-0810

that the wavelet ψ has M vanishing moments and M continuous derivatives ($M \geqslant \gamma$) that have a fast decay. This means that for any $0 \leqslant k \leqslant M$ and $m \in \mathbf{N}$ there exists a constant C_m such that for all $x \in \mathbf{R}$

$$\left|\psi^{(k)}(x)\right| \leqslant \frac{C_m}{1+|x|^m}.$$

Then it is known [9] that there exists $C_f > 0$ such that

$$|\langle f, \psi_{j,k}\rangle| \leqslant \frac{C_f}{2^{j(\gamma+1/2)}}. \tag{2}$$

It is further assumed that the wavelet ψ satisfies these requirements.

Note. If we additionally assume that ψ has a compact support, then the uniform Lipschitz regularity requirement may be replaced by a piecewise Lipschitz regularity [9]. The results of this paper will not change.

In practice the signal function f is defined by its samples. Suppose that the number of these samples is 2^J for some $J > 0$. The discrete wavelet transform is a multiplication of the vector of values of f by the orthogonal matrix defined by the wavelet function ψ. Due to the orthogonality of the matrix the discrete wavelet coefficients are related to the continuous ones as $\mu_{j,k} = 2^{J/2}\langle f, \psi_{j,k}\rangle$ [9]. In addition the observations always contain some noise. In this paper we consider the following data model:

$$X_i = f_i + w_i, \qquad i = 1, \dots, 2^J,$$

where f_i are clean samples of the signal function and $w_i \sim N(0, \sigma^2)$ are samples from a white Gaussian noise. Applying the discrete wavelet transform we obtain the following model of the empirical wavelet coefficients:

$$Y_{j,k} = \mu_{j,k} + W_{j,k}, \;\; j = 0, \dots, J-1, \;\; k = 0, \dots, 2^j - 1,$$

where $W_{j,k}$ have the same statistical structure as w_i.

To remove the noise one usually uses a thresholding procedure. This procedure removes small coefficients which are considered to be pure noise.

Denote by $\hat{Y}_{j,k}$ the estimate of the wavelet coefficient which is obtained with the use of thresholding function $\rho_T(x)$ and the threshold value T: $\hat{Y}_{j,k} = \rho_T(Y_{j,k})$. We will consider the soft thresholding $\rho_T^{(s)}(x) = \operatorname{sign}(x)(|x| - T)_+$ and the hard thresholding $\rho_T^{(h)}(x) = x\mathbb{I}(|x| > T)$.

Consider the cost function $r_J(f)$ defined for a given critical value $\varepsilon > 0$ as

$$r_J(f) = \mathsf{P}\left(\max_{j,k}\left|\hat{Y}_{j,k} - \mu_{j,k}\right| > \varepsilon\right). \tag{3}$$

Note that as the number of samples grows, $r_J(f)$ tends to 1. The goal of this paper is to find an optimal threshold ensuring the minimum loss in the sense that the rate of convergence of $r_J(f)$ to 1 is the slowest. Due to the relation

$$1 - \mathsf{P}\left(\max_{j,k}\left|\hat{Y}_{j,k} - \mu_{j,k}\right| > \varepsilon\right) = \prod_{j=0}^{J-1}\prod_{k=0}^{2^j-1}\mathsf{P}\left(\left|\hat{Y}_{j,k} - \mu_{j,k}\right| \leqslant \varepsilon\right)$$

this problem is equivalent to finding the threshold for which the rate of convergence of the sum

$$S_J(f) = \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1}\log\mathsf{P}\left(\left|\hat{Y}_{j,k} - \mu_{j,k}\right| \leqslant \varepsilon\right) \tag{4}$$

to $-\infty$ is the slowest.

The main results are formulated in the minimax framework for the class of functions $f \in \mathrm{Lip}(\gamma, L)$, where the cost function is defined as

$$R_J = \sup_{f \in \mathrm{Lip}(\gamma,L)} r_J(f). \tag{5}$$

Note that the maximum reasonable threshold is the so-called universal threshold $T_U = \sigma\sqrt{2\log 2^J}$. It follows from the relation

$$\lim_{J\to\infty} \mathsf{P}\left(T_U - \frac{\sigma \log\log 2^J}{\sqrt{\log 2^J}} \leqslant \max_{j,k} |W_{j,k}| \leqslant T_U\right) = 1.$$

This threshold with a high probability removes most of the noise, while the higher threshold may damage the useful signal components [9]. Hence, we assume that $T \leqslant T_U$. Note also that the "reasonable" threshold T should grow with J [2].

By the symbol $\asymp$ we denote the order of the considered value, i.e., $a_J \asymp b_J$, if there exists a positive constant C such that $\lim\limits_{J\to\infty} a_J/b_J = C$.

3. The optimal threshold for the soft thresholding

Let the estimates of the wavelet coefficients be obtained by the soft thresholding: $\hat{Y}_{j,k} = \rho_T^{(s)}(Y_{j,k})$.

Let the function $g_1(J) > 0$ decrease arbitrarily slowly to zero and let the function $g_2(J) > 0$ grow to infinity so that

$$\log g_2(J) = o(\sqrt{\log 2^J}), \qquad J \to \infty. \tag{6}$$

The inequality (2) allows splitting the set of indices $\{0, \ldots, J-1\}$ into three classes depending on the values of $|\mu_{j,k}|$. Let j_1 and j_2 $(j_1 < j_2)$ be such that

$$|\mu_{j,k}| \leqslant C(g_1(J))^{-(\gamma+1/2)}, \qquad j_1 \leqslant j \leqslant j_2 - 1,$$

$$|\mu_{j,k}| \leqslant C(g_2(J))^{-(\gamma+1/2)}, \qquad j_2 \leqslant j \leqslant J - 1.$$

From (2) we have

$$j_i = \frac{J}{2\gamma+1} + \log_2 g_i(J) + C, \quad i = 1, 2. \tag{7}$$

Let us split (4) into three sums:

$$\begin{aligned} S_J(f) = \sum_{j=0}^{j_1-1}\sum_{k=0}^{2^j-1} \log\mathsf{P}\left(\left|\hat{Y}_{j,k} - \mu_{j,k}\right| \leqslant \varepsilon\right) + \\ + \sum_{j=j_1}^{j_2-1}\sum_{k=0}^{2^j-1} \log\mathsf{P}\left(\left|\hat{Y}_{j,k} - \mu_{j,k}\right| \leqslant \varepsilon\right) + \sum_{j=j_2}^{J-1}\sum_{k=0}^{2^j-1} \log\mathsf{P}\left(\left|\hat{Y}_{j,k} - \mu_{j,k}\right| \leqslant \varepsilon\right) \equiv \\ \equiv S_1 + S_2 + S_3. \end{aligned} \tag{8}$$

Consider S_3. Note that for any fixed $\varepsilon > 0$ there exists such $J_0 = J_0(\varepsilon)$ that $C(g_2(J))^{-(\gamma+1/2)} \leqslant \varepsilon$ for all $J > J_0$. Thus we have $|\mu_{j,k}| \leqslant \varepsilon$ for $j_2 \leqslant j \leqslant J-1$. Hence, for any summand from S_3 we have for $J > J_0$:

$$\begin{aligned} \log\mathsf{P}\left(\left|\hat{Y}_{j,k} - \mu_{j,k}\right| \leqslant \varepsilon\right) = \log\Big[\mathsf{P}(|\mu_{j,k}| \leqslant \varepsilon, |Y_{j,k}| \leqslant T) + \\ + \mathsf{P}(|Y_{j,k} - \mu_{j,k} - T| \leqslant \varepsilon, Y_{j,k} > T) + \mathsf{P}(|Y_{j,k} - \mu_{j,k} + T| \leqslant \varepsilon, Y_{j,k} < -T)\Big] = \end{aligned}$$

$$= \log\left[2\Phi\left(\frac{T+\varepsilon}{\sigma}\right) - 1\right].$$

Using the relations

$$1-\Phi(x) \asymp \frac{\phi(x)}{x}, \qquad x\to\infty,$$

$$\log(1-x) \asymp -x, \qquad x \to 0,$$

we conclude that

$$S_3 \asymp -2^J \frac{\exp\left\{-\frac{T^2}{2\sigma^2}\right\}\exp\left\{-\frac{T\varepsilon}{\sigma^2}\right\}}{T}. \tag{9}$$

In order to find the lower estimate for the optimal threshold, note that the sum $S_J(f)$ from (4) tends to $-\infty$ the faster, the more samples of the signal function satisfy the inequality $|\mu_{j,k}| > \varepsilon$. According to the definition (7) the maximum number of such samples has the order of 2^{j_2}. Assuming that all the summands in S_1 and S_2 from (8) have $|\mu_{j,k}| > \varepsilon$, we conclude that

$$S_1 + S_2 \asymp \sum_{j=0}^{j_2-1}\sum_{k=0}^{2^j-1}\log\Big[\mathsf{P}(T-\varepsilon \leqslant Y_{j,k}-\mu_{j,k} \leqslant T+\varepsilon, Y_{j,k} > T) +$$

$$+\mathsf{P}(-T-\varepsilon \leqslant Y_{j,k}-\mu_{j,k} \leqslant -T+\varepsilon, Y_{j,k} < -T)\Big] =$$

$$= \sum_{j=0}^{j_2-1}\sum_{k=0}^{2^j-1}\log\left[\Phi\left(\frac{T+\varepsilon}{\sigma}\right) - \Phi\left(\frac{T-\varepsilon}{\sigma}\right)\right] \asymp$$

$$\asymp 2^{\frac{J}{2\gamma+1}} g_2(J) \log \frac{\exp\left\{-\frac{T^2}{2\sigma^2}\right\}}{T} \asymp -2^{\frac{J}{2\gamma+1}} g_2(J) T^2.$$

Now let us equate the orders of $S_1 + S_2$ and S_3. Let

$$T_*^{(s)} = \sigma\sqrt{\frac{4\gamma}{2\gamma+1}\log 2^J} - \varepsilon, \tag{10}$$

$$T_i^{(s)} = \sigma\sqrt{\frac{2\gamma+1}{4\gamma}} \cdot \frac{\log\left((\log 2^J)^{3/2} g_i(J)\right)}{\sqrt{\log 2^J}}, \qquad i = 1, 2. \tag{11}$$

It is easy to see that the equality of the orders is attained with the threshold

$$T_m^{(s)} = T_*^{(s)} - T_2^{(s)}, \tag{12}$$

which thereby is the lower estimate (up to the $O(1/\sqrt{\log 2^J})$ terms) for the optimal threshold in the cost function R_J, since any lower threshold would increase the rate of convergence of R_J to 1.

Let us now find the upper estimate for the optimal threshold. Note that for any C_f from (2) there exists a function $f \in \mathrm{Lip}(\gamma, L)$ such that the inequality (2) becomes the equality for $0 \leqslant j \leqslant j_1 - 1$ [9]. Hence, since $T \leqslant T_U$, there exists such $J_1 > 0$ that for all $\varepsilon > 0$ and $J > J_1$ we have $|\mu_{j,k}| > \varepsilon$ when $0 \leqslant j \leqslant j_1 - 1$. Repeating the above considerations we conclude that

$$S_1 \asymp -2^{\frac{J}{2\gamma+1}} g_1(J) T^2.$$

Let us equate the orders of S_1 and S_3. In this case the threshold equals

$$T_M^{(s)} = T_*^{(s)} - T_1^{(s)}, \tag{13}$$

where $T_*^{(s)}$ and $T_1^{(s)}$ are defined in (10) and (11) respectively.

Note that we omitted the sum S_2 in these arguments. This means that the real value of the optimal threshold T should be less than (13), since by reducing the value of T the order of the sum S_3 can be increased to its real value (taking into account the sum S_2). Note also that the value of the threshold should be the lowest possible for a given order of the cost function, since the higher threshold could damage the important components of the signal function.

The above considerations lead to the following statement.

Theorem. *For the optimal value of the soft threshold minimizing the rate of convergence of the cost function* (5) *to* 1, *the following inequalities hold starting with some* J :

$$T_m^{(s)} \leqslant T \leqslant T_M^{(s)},$$

where $T_m^{(s)}$ *and* $T_M^{(s)}$ *are defined in* (12) *and* (13) *respectively.*

Note. The thresholds defined in (12) and (13) have the same growing component $T_*^{(s)}$, and since (6) holds, the difference $|T_m^{(s)} - T_M^{(s)}|$ tends to zero. This means that the real value of the optimal soft threshold minimizing the cost function also has the main component $T_*^{(s)}$. Note also that the growth rate of $T_*^{(s)}$ coincides with that of asymptotically optimal threshold minimizing the mean squared risk in the problem of estimating the Lipschitz function by thresholding its wavelet coefficients [2].

4. Hard thresholding

Let the estimates of the wavelet coefficients be obtained by the hard thresholding: $\hat{Y}_{j,k} = \rho_T^{(h)}(Y_{j,k})$.

Consider the cost function (5). Note that for an arbitrary $\varepsilon > 0$ there exists a function $f \in \mathrm{Lip}(\gamma, L)$ such that the inequality (2) becomes the equality for some j_h and $|\mu_{j_h,k}| > \varepsilon$. Hence there exists such $J_2 > 0$ that for $J > J_2$ we have $|\mu_{j_h,k}| > \varepsilon$, $T - \mu_{j_h,k} > \varepsilon$ and $-T - \mu_{j_h,k} < -\varepsilon$. In this case

$$\mathsf{P}\left(\left|\hat{Y}_{j_h,k} - \mu_{j_h,k}\right| \leqslant \varepsilon\right) = \mathsf{P}(-\varepsilon \leqslant Y_{j_h,k} - \mu_{j_h,k} \leqslant \varepsilon, Y_{j_h,k} - \mu_{j_h,k} > T - \mu_{j_h,k}) +$$

$$+\mathsf{P}(-\varepsilon \leqslant Y_{j_h,k} - \mu_{j_h,k} \leqslant \varepsilon, Y_{j_h,k} - \mu_{j_h,k} < -T - \mu_{j_h,k}) = 0,$$

and thus for $J > J_2$

$$R_J = \sup_{f \in \mathrm{Lip}(\gamma,L)} \mathsf{P}\left(\max_{j,k}\left|\hat{Y}_{j,k} - \mu_{j,k}\right| > \varepsilon\right) = 1.$$

This fact means that there is no sense in estimating the cost function (5) for the hard thresholding method. It is an interesting observation since this is not the case when estimating the mean squared risk.

Acknowledgments

This research is partly supported by the Russian Foundation for Basic Research (project No. 16–07–00736).

REFERENCES

1. D. Donoho and I. Johnstone, "Minimax estimation via wavelet shrinkage," *Ann. Stat.*, **26**, No. 3, 879–921 (1998).
2. M. Jansen, *Noise Reduction by Wavelet Thresholding*, Springer, Berlin (2001).
3. D. Donoho and I. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Am. Stat. Assoc.*, **90**, 1200–1224 (1995).
4. A.M. Altaher and M.T. Ismail, "A comparison of some thresholding selection methods for wavelet regression," *World Acad. Sci. Eng. Tech.*, **62**, 119–125 (2010).

5. A.A. Kudryavtsev and O.V. Shestakov, "Asymptotic behavior of the threshold minimizing the average probability of error in calculation of wavelet coefficients," *Dokl. Math.*, **93**, No. 3, 295–299 (2016).
6. A.V. Markin and O.V. Shestakov, "Consistency of risk estimation with thresholding of wavelet coefficients," *Moscow Univ. Comput. Math. Cybern.*, **34**, No. 1, 22–30 (2010).
7. O.V. Shestakov, "Asymptotic normality of adaptive wavelet thresholding risk estimation," *Dokl. Math.*, **86**, No. 1, 556–558 (2012).
8. J. Sadasivan, S. Mukherjee, and C.S. Seelamantula, "An optimum shrinkage estimator based on minimum-probability-of-error criterion and application to signal denoising," in: *Proc. IEEE ICASSP* (2014), pp. 4249–4253.
9. Mallat S., *A Wavelet Tour of Signal Processing*, Academic Press, New York (1999).