element of $\boldsymbol{D}$ is equal to 0 for $n+1 \leq k \leq m$. Consequently, $\overline{a}_k = 0$ w.p. 1 for $n+1 \leq k \leq m$. Therefore, we conclude that $\hat{\overline{\boldsymbol{W}}}_s$ is block diagonal. The upper left $n \times n$ block is a diagonal matrix, with diagonal elements $c/\sqrt{d_k}$; the lower right block is arbitrary, since $\overline{b}_k = \overline{a}_k = 0$ regardless of the choice of this block. We, therefore, choose $\hat{\overline{\boldsymbol{W}}}_s$ to be a diagonal matrix with the first n diagonal elements equal to $c/\sqrt{d_k}$ and the remaining diagonal elements equal to 0. Thus, $\hat{\overline{\boldsymbol{W}}}_s = c(\boldsymbol{D}^{1/2})^{\dagger}$, and

$$\hat{\boldsymbol{W}}_s = c\boldsymbol{V}(\boldsymbol{D}^{1/2})^{\dagger}\boldsymbol{V}^* = c(\boldsymbol{C}_a^{1/2})^{\dagger}. \tag{36}$$

If we choose to minimize the MSE with respect to c as well, then it is straightforward to show that the optimal value of c is given by

$$\alpha_s = \frac{1}{n}\sum_{k=1}^{n}\sqrt{d_k}.$$

REFERENCES

[1] E. M. Friel and K. M. Pasala, "Direction finding with compensation for a near field scatterer," in *Proc. Int. Symp. Antennas and Propagation Society*, 1995, pp. 106–109.

[2] R. J. Piechocki, N. Canagarajah, and J. P. McGeehan, "Improving the direction-of-arrival resolution via double code filtering in WCDMA," in *Proc. 1st Int. Conf. 3G Mobile Communication Technologies*, Mar. 2000, pp. 204–207.

[3] Y. C. Eldar, A. V. Oppenheim, and D. Egnor, "Orthogonal and projected orthogonal matched filter detection," *Signal Processing*, submitted for publication.

[4] Y. C. Eldar, "Quantum signal processing," Ph.D. dissertation, MIT, Cambridge, MA, Dec. 2001. also available at http://allegro.mit.edu/dspg/publications/TechRep/index.html.

[5] Y. C. Eldar and A. M. Chan, "An optimal whitening approach to linear multiuser detection," *IEEE Trans. Inform. Theory*, to be published.

[6] Y. C. Eldar and G. D. Forney, Jr., "On quantum detection and the square-root measurement," *IEEE Trans. Inform. Theory*, vol. 47, pp. 858–872, Mar. 2001.

[7] Y. C. Eldar, "Least-squares inner product shaping," *Linear Alg. Appl.*, vol. 348, pp. 153–174, May 2002.

[8] ——, "Least-squares orthogonalization using semidefinite programming," *Linear Alg. Appl.*, submitted for publication.

[9] Y. C. Eldar and G. D. Forney, Jr., "Optimal tight frames and quantum measurement," *IEEE Trans. Inform. Theory*, vol. 48, pp. 599–610, Mar. 2002.

[10] Y. C. Eldar and H. Bölcskei, "Geometrically uniform frames," *IEEE Trans. Inform. Theory*, vol. 49, pp. 993–1006, Apr. 2003.

[11] Y. C. Eldar, "Minimum mean-squared error covariance shaping," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, vol. 6, Hong Kong, Apr. 2003, pp. 713–716.

[12] Y. C. Eldar and A. V. Oppenheim, "Covariance shaping least-squares estimation," *IEEE Trans. Signal Processing*, vol. 51, pp. 696–697, Mar. 2003.

[13] Y. C. Eldar, "Covariance shaping approach to linear least-squares estimation," in *Proc. Asilomar Conf. Signals, Systems, and Computers*, Nov. 2002.

[14] Y. C. Eldar and A. V. Oppenheim, "Quantum signal processing," *IEEE Signal Processing Mag.*, pp. 12–32, Nov. 2002.

[15] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: Johns Hopkins Univ. Press, 1996.

[16] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

[17] J. J. Atick and A. N. Redlich, "Convergent algorithm for sensory receptive field development," *Neurak Comp.*, vol. 5, pp. 45–60, 1993.

[18] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1992.

[19] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ: Prentice-Hall, 1993.

Deviation Bounds for Wavelet Shrinkage

Dawei Hong and Jean-Camille Birget

***Abstract*—We analyze the wavelet shrinkage algorithm of Donoho and Johnstone in order to assess the quality of the reconstruction of a signal obtained from noisy samples. We give a deviation estimate for the maximum squared error (and, consequently, for the average squared error), under the assumption that the signal comes from a Hölder class, and the noise samples are independent, of zero mean, and bounded. Our main technique is Talagrand's isoperimetric theorem. Our result shows a better behavior of the wavelet shrinkage.**

***Index Terms*—Deviation bound, maximum squared error, wavelet shrinkage.**

I. INTRODUCTION

We address the classical problem of the reconstruction of signal samples from noisy samples. We consider an original signal of bounded duration $f: t \in [0, 1] \to f(t) \in \mathrm{R}$. We also have additive noise $e: [0, 1] \to \mathrm{R}$. Thus, the observed *noisy signal* at time t is $y(t) = f(t) + e(t)$.

We sample the noisy signal at n uniformly spaced instants and we denote the sample values by

$$y_i = f_i + e_i = f\left(\frac{i}{n}\right) + e\left(\frac{i}{n}\right) \qquad (\text{for} 1 \leq i \leq n).$$

Our goal is to recover a good approximation of the original signal samples $(f_1, \ldots, f_n)$ from the noisy signal samples $(y_1, \ldots, y_n)$. For this to be possible we need some assumptions that distinguish the signal from the noise.

- The original signal f has a certain degree of "smoothness," i.e., f belongs to a Hölder class $\Lambda^{\alpha}(M)$ for some $\alpha > 0$ and $M > 0$.
- The noise is "random," i.e., $(e_1, \ldots, e_n)$ consists of n independent random variables.

The *Hölder classes* are defined as follows:

$$\text{For } 0 < \alpha \leq 1, \quad \Lambda^{\alpha}(M) = \{h \in \mathrm{R}^{[0,1]} : (\forall x_1, x_2 \in [0,1]), |h(x_1) - h(x_2)| \leq M|x_1 - x_2|^{\alpha}\}.$$

$$\text{For } 1 < \alpha, \quad \Lambda^{\alpha}(M) = \{h \in \mathrm{R}^{[0,1]} : (\forall x \in [0,1]) \; |h'(x)| \leq M, \; h^{\lfloor \alpha \rfloor} \text{ exists, and } (\forall x_1, x_2 \in [0,1]) \; |h^{\lfloor \alpha \rfloor}(x_1) - h^{\lfloor \alpha \rfloor}(x_2)| \leq M|x_1 - x_2|^{\alpha - \lfloor \alpha \rfloor}\}.$$

Let $(\tilde{y}_1, \ldots, \tilde{y}_n)$ be an approximation of $(f_1, \ldots, f_n)$, obtained from $(y_1, \ldots, y_n)$. Most commonly, the closeness of this approximation is measured by $\frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - f_i)^2$ or by the expectation $\mathbf{E}[\frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - f_i)^2]$.

The wavelet shrinkage algorithm of Donoho and Johnstone [6], [7] is a very efficient tool for finding good estimates $\tilde{y}$. In outline, the algorithm works as follows:

Manuscript received May 14, 2001; revised February 24, 2003. The work of J.-C. Birget was supported in part by the National Science Foundation under Grant DMS-9970471.

The authors are with the Department of Computer Science, Rutgers University at Camden, Camden, NJ 08102 USA (e-mail: dhong@camden.rutgers.edu; birget@camden.rutgers.edu).

Communicated by J. A. O'Sullivan, Associate Editor for Detection and Estimation.

Digital Object Identifier 10.1109/TIT.2003.813482

(Step 0) Choose a wavelet system with N vanishing moments ($N \geq \alpha$); choose a level of coarseness $J_0 \geq 0$ (J_0 will depend on α), and consider the multiresolution chain of Hilbert spaces $V_{J_0} \subset V_{J_0+1} \subset \cdots \subset V_j \subset \cdots$.

(Step 1) Apply the *discrete wavelet transform* (DWT) to the noisy signal samples $(y_1, \ldots, y_n)$, where $n \geq 2^{J_0}$. This yields the "empirical wavelet coefficients" $(\xi_1, \ldots, \xi_n)$.

(Step 2) Fix a "threshold" $\lambda_n (>0)$ and apply either "hard" or "soft thresholding" to $(\xi_1, \ldots, \xi_n)$.

Hard thresholding consists of replacing each ξ_i by 0 when $|\xi_i| \leq \lambda_n$, and keeping ξ_i unchanged when $|\xi_i| > \lambda_n$.

Soft thresholding consists of transforming each ξ_i as follows: ξ_i is replaced by 0 if $|\xi_i| \leq \lambda_n$; if $\xi_i > \lambda_n$, ξ_i is replaced by $\xi_i - \lambda_n$; if $\xi_i < -\lambda_n$, ξ_i is replaced by $\xi_i + \lambda_n$.

(Step 3) Apply the inverse DWT to the result of (2). This yields the estimate $(\tilde{y}_1, \ldots, \tilde{y}_n)$.

To what extent does wavelet shrinkage depend on the smoothness conditions of the signal f and on the randomness conditions of the noise samples e_i, and how do the estimators $\tilde{y}_i$ approximate the original signal f? In [6], [7], it was assumed that the e_i are independent and identically distribute (i.i.d.) Gaussian variables with distribution $N(0, \sigma^2)$, and the threshold was chosen to be

$$\lambda_n = \sigma\sqrt{2\frac{\log n}{n}}.$$

Assuming that $f \in \Lambda^\alpha(M)$ (the Hölder class) with $\alpha > 0$, it is proved in [6], [7] that

$$\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - f_i)^2\right] < C \cdot \left(\frac{1}{n}\log n\right)^{\frac{2\alpha}{1+2\alpha}}$$

where C depends only on M and on the wavelet system used. It was observed in [6], [7] (the proofs are due to Lepskii [9] and to Brown and Low [3]) that this upper bound is optimal over all possible algorithms, if the parameters α and M are *not* known. For the optimality of the wavelet shrinkage algorithm it is important that the threshold be of the form $c \cdot \sqrt{\frac{\log n}{n}}$ (where c does not depend on n).

Since the publication of [6], [7] there has been further progress on wavelet shrinkage ([13, Ch. 6] is an excellent reference up to 1999). Most recently, Averkamp and Houdré [1], [2] expanded the scope of wavelet shrinkage by allowing the noise samples e_i to have different distributions F_i, chosen from a wide class of distributions. They show in [1, p. 32] that the error expectation of the wavelet shrinkage algorithm for bounded noise is roughly the same as for Gaussian noise, if the parameters α and M of the Hölder class of the signal are not known. They also discuss various choices of thresholds.

All the results on wavelet shrinkage in the literature so far have evaluated the quality of the approximation by bounding the expectation $\mathbf{E}[\frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - f_i)^2]$. In this correspondence, we study the deviation bound (rather than just the expectation) of the maximum squared error $\max\{(\tilde{y}_i - f_i)^2 : 1 \leq i \leq n\}$.

Assumptions: We assume that the signal f belongs to a Hölder class $\Lambda^\alpha(M)$, and that the noise samples e_i are independent random variables (with possibly different distributions). The only restrictions on the distributions are that they have bounded support (contained in the interval $[-\frac{b}{2}, \frac{b}{2}]$), and zero mean.

Abbreviations: We denote by $Q_{\mathrm{avg}}(n)$ the average squared error $\frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - f_i)^2$, and by $Q_{\max}(n)$ the maximum squared error $\max\{(\tilde{y}_i - f_i)^2 : 1 \leq i \leq n\}$. Notice that both $Q_{\mathrm{avg}}(n)$ and $Q_{\max}(n)$ are random variables.

The main result of this correspondence is the following deviation estimate. Let

$$\lambda_{n,\delta} = C_\varphi b\left(1 + 2\sqrt{(1+\delta)\ln 2}\right)\sqrt{\frac{\log n}{n}}$$

where C_φ depends only on the wavelet system being used.

Theorem 1.1: For wavelet shrinkage with threshold $\lambda_{n,\delta}$ we have for all $n \geq n_0$ and all $\delta > o$

$$P\left(Q_{\max}(n) \leq (c_1 + c_2\delta)\left(\frac{\log n}{n}\right)^{\frac{2\alpha}{1+2\alpha}}\right) \geq 1 - \frac{9}{n^{1+\delta}}$$

where c_1 and c_2 depend only on b, M, and α. Consequently

$$P\left(Q_{\mathrm{avg}}(n) \leq (c_1 + c_2\delta)\left(\frac{\log n}{n}\right)^{\frac{2\alpha}{1+2\alpha}}\right) \geq 1 - \frac{9}{n^{1+\delta}}.$$

A theoretical lower bound for the minimum number of samples n_0 is $2^9 = 512$ when $0 < \alpha \leq 1$; when $\alpha > 1$, $n_0 = (4\alpha+2)^{2\alpha+2} \cdot (\log_2(4\alpha+2))^2$. In most practical applications, it is the case where $\alpha \leq 1$. Taking more than 512 samples from a signal is quite easy for today's computer.

Many applications of denoising are carried out in real-time environment. For example, recognizing a warship by its shape, ours or enemy's, in foggy sea. A signal is presented a few times. The computer system must respond promptly. Suppose that the wavelet shrinkage is employed for denoising. The sharper the deviation estimate, the more we can trust the outputs from the wavelet shrinkage. We prove a deviation bound on the maximum squared error that is stronger than the usual expectation error bound given in analysis of the wavelet shrinkage. Theorem 1.1 shows a better behavior for the maximum squared error and thus from application viewpoint gives more confidence in the wavelet shrinkage.

The main technique we use to prove Theorem 1.1 is Talagrand's isoperimetric theorem [12], which has been very successfully applied to probabilistic analysis of combinatorial optimizations [11], but has not been used in analysis of the wavelet shrinkage. We find that the discrete wavelet transform can fit into the framework of Talagrand's convex distance (it is technically nontrivial), which gives a new approach to estimates for the wavelet shrinkage.

The remainder of this correspondence is organized as follows: In the next section, we give preliminaries on wavelets and Talagrand's theorem. Then, in the last section, we prove Theorem 1.1.

II. Preliminaries

A. Wavelets

We will usually follow the notation of [5] regarding wavelets, the only exception being that we reverse the multiresolution indexes. Moreover, we only consider real-valued functions with domain $[0, 1]$. So we have a sequence of real Hilbert spaces $V_{J_0} \subset V_{J_0+1} \subset \cdots \subset V_j \subset \ldots$ such that the closure of $\bigcup_j V_j$ is $\mathrm{L}^2[0, 1]$. We let $V_{j+1} = V_j \oplus W_j$ (orthogonal complement). Since we are in the case of compactly supported functions each V_j is a finite-dimensional real vector space (of dimension 2^j), with orthonormal basis $\{\varphi_{j,k} : 0 \leq k \leq 2^j - 1\}$, derived from a scaling function φ. Let ψ be the wavelet function corresponding to φ, and let $\{\psi_{j,k} : 0 \leq k \leq 2^j - 1\}$ be the corresponding orthonormal basis of W_j.

For any function $g \in \mathrm{L}^2[0, 1]$, we define the piecewise-constant function $\overline{g}$: $[0, 1] \to \mathrm{R}$ as follows: $\overline{g}(x) = g(\frac{k}{n})(=g_k)$ if $\frac{k-1}{n} < x \leq \frac{k}{n}$ for some $k = 1, \ldots, n$; $\overline{g}(x) = 0$ if $x \notin]0, 1]$. The discrete wavelet transform of a vector $(g_1, \ldots, g_n)$ can be obtained by taking

the wavelet coefficients of the piecewise-constant function $\overline{g}$. These *wavelet coefficients* are

$$c_{j,k}^{(g)} = \langle \overline{g},\ \varphi_{j,k} \rangle = \int_0^1 \overline{g}(x)\varphi_{j,k}(x)\,dx$$

and

$$d_{j,k}^{(g)} = \langle \overline{g},\ \psi_{j,k} \rangle = \int_0^1 \overline{g}(x)\psi_{j,k}(x)\,dx.$$

Then, for any integer $J \geq J_0$, we have almost everywhere

$$\overline{g}(x) = \sum_{k=0}^{2^J-1} c_{J,k}^{(g)}\varphi_{J,k}(x) + \sum_{j=J}^{+\infty}\sum_{k=0}^{2^j-1} d_{j,k}^{(g)}\psi_{j,k}(x).$$

In this correspondence, we will use two wavelet systems. The Haar wavelets (because of their simplicity, especially for programming purposes), and the interval wavelets with predefined vanishing moments, based on Daubechies wavelets (Cohen, Daubechies, Jawerth, Vial [4]).

For the *Haar wavelets*, the scaling function is $\varphi(x) = 1$ when $0 < x \leq 1$, and $\varphi(x) = 0$ otherwise. Hence, $\varphi_{j,k}(x) = 2^{j/2}$ when $k2^{-j} < x \leq (k+1)2^{-j}$, and $\varphi_{j,k}(x) = 0$ otherwise. The Haar wavelet function is $\psi(x) = 1$ if $0 < x \leq \frac{1}{2}$, $\psi(x) = -1$ if $\frac{1}{2} < x \leq 1$, and $\psi(x) = 0$ otherwise. Hence, $\psi_{j,k}(x) = 2^{j/2}$ if $k2^{-j} < x \leq (k+\frac{1}{2})2^{-j}$, $\psi_{j,k}(x) = -2^{j/2}$ if $(k+\frac{1}{2})2^{-j} < x \leq (k+1)2^{-j}$, and $\psi_{j,k}(x) = 0$ otherwise.

For the *interval wavelet system* of [4], with N *vanishing moments*, the scaling function φ and the wavelet function ψ are complicated. However, all we need to know about them is the following.

- A multiresolution of $L^2[0,\ 1]$ is obtained, with an orthonormal basis for V_j when $j > J_0$

$$\{\varphi_{j,k}\colon 1 \leq k < 2^j - 2N\} \cup \{\varphi_{j,i}^{\text{left}},\ \varphi_{j,i}^{\text{right}}\colon 0 \leq i < N\}.$$

Each $\varphi_{j,k}$ has support $[k2^{-j},\ (2N-1+k)2^{-j}]$, each $\varphi_{j,i}^{\text{left}}$ has support $[0,\ i2^{-j}]$, and each $\varphi_{j,i}^{\text{right}}$ has support $[1 - i2^{-j},\ 1]$.

The decomposition level J_0 is chosen so that $J_0 \geq 1 + \log_2(2N - 1)$. For signals in the Hölder class $\Lambda^\alpha(M)$ we require the number of vanishing moments to be $N \geq \alpha$.

- We also have an orthonormal basis for W_j

$$\{\psi_{j,k}\colon 1 \leq k < 2^j - 2N\} \cup \{\psi_{j,i}^{\text{left}},\ \psi_{j,i}^{\text{right}}\colon 0 \leq i < N\}$$

with the same supports as the corresponding φ functions.

- φ and ψ are bounded on $[0,\ 1]$ by a constant $C > 0$, independent of x and N: $\forall\, x \in [0,\ 1]$, $|\varphi(x)|, |\psi(x)| \leq C$.

For $0 \leq k < 2^j - 2N$ ("inside the interval"), $\varphi_{j,k}(x) = 2^{j/2}\varphi(2^j x - k)$.

At the ends of the interval $[0,\ 1]$ we have for $0 \leq i < N$, (see [4])

$$\varphi_{j,i}^{\text{left}}(x) = \sum_{h=1}^{2N-1} (-h)^i \varphi(2^j x + h).$$

A similar formula holds on the right end of the interval $[0,\ 1]$.

Assuming that n is a power of 2, $n = 2^J$, we have for the function $\overline{y}$, relative to any wavelet system

$$\overline{y}(x) = \sum_{k=0}^{2^J-1} \langle \overline{y},\ \varphi_{J,k}\rangle \varphi_{J,k}(x).$$

Thus, for any J_1 with $0 \leq J_1 < J$, the DWT transforms $(y_1,\ \ldots,\ y_n)$ to

$$\sqrt{n}\ \Big(c_{J_1,0}^{(\overline{y})},\ \ldots,\ c_{J_1,2^J-1}^{(\overline{y})},\ d_{J_1,0}^{(\overline{y})},\ \ldots,\ d_{J_1,2^J-1}^{(\overline{y})},\ \ldots,\ d_{J-1,0}^{(\overline{y})},\ \ldots,\ d_{J-1,2^J-1}^{(\overline{y})}\Big).$$

The DWT is an orthogonal transformation (represented by an orthogonal matrix W).

We will always assume that n is a power of 2: $n = 2^J$. Throughout this correspondence, $\log$ will refer to $\log_2$, and $\ln$ will denote the natural logarithm.

Let us now return to the analysis of a noisy signal $y(t) = f(t) + e(t)$.

Proposition 2.1: With respect to the Haar wavelets, the wavelet coefficients of the function e have the following properties.

(H1) For all $j \in [0,\ 2^J]$ and all $k \in [0,\ 2^{j-1} - 1]$

$$c_{j,k}^{(e)} = 2^{-J+j/2} \sum_{i=0}^{2^{J-j}-1} e_{i+1+k2^{J-j}}.$$

(H2) For all j and k as in (H1)

$$d_{j,k}^{(e)} = 2^{-J+j/2} \sum_{i=0}^{2^{J-j-1}-1} \left(e_{i+1+k2^{J-j}} - e_{i+1+(k+\frac{1}{2})2^{J-j}}\right).$$

For any function $f\colon [0,\ 1] \to \mathrm{R}$ belonging to $\Lambda^{(\alpha)}(M)$ with $0 < \alpha \leq 1$ we have the following.

(H3) For all $j \in [0,\ 2^J]$ and all $k \in [0,\ 2^{j-1} - 1]$

$$\left|d_{j,k}^{(f)}\right| < M2^{-j(\frac{1}{2}+\alpha)}. \qquad \square$$

Proposition 2.2: With respect to the interval wavelet system [4], the wavelet coefficients of the function e have the following properties.

(D1) For all $j \in [0,\ 2^J]$ and all $k \in [0,\ 2^{j-1} - 1]$

$$c_{j,k}^{(e)} = 2^{-J+j/2} \sum_{i=0}^{2^{J-j}-1} \alpha_{i,j,k} e_{i+1+k2^{J-j}}$$

for some numbers $\alpha_{i,j,k}$ that do not depend on the noise function e. Moreover, $|\alpha_{i,j,k}| < C_\varphi$ for some constant $C_\varphi \geq 1$ depending only on the wavelet system.

(D2) For all j and k as in (D1)

$$d_{j,k}^{(e)} = 2^{-J+j/2} \sum_{i=0}^{2^{J-j}-1} \beta_{i,j,k} e_{i+1+k2^{J-j}}$$

for some numbers $\beta_{i,j,k}$ that do not depend on the noise function e. Moreover, $|\beta_{i,j,k}| < C_\varphi$ where $C_\varphi \geq 1$ depends only on the wavelet system.

Suppose $f\colon [0,\ 1] \to \mathrm{R}$ belongs to $\Lambda^{(\alpha)}(M)$ with $1 < \alpha$, and suppose the number of vanishing moments N of the wavelet system satisfies $N \geq \alpha$. Then we have the following.

(D3) For all $j \in [0,\ 2^J]$ and all $k \in [0,\ 2^{j-1} - 1]$

$$\left|d_{j,k}^{(f)}\right| < C_\varphi M 2^{-j(\frac{1}{2}+\alpha)}$$

where $C_\varphi \geq 1$ depends only on the wavelet system. $\square$

The two propositions in this subsection can be proved by straightforward calculations. We omit the proofs.

B. Talagrand's Isoperimetric Theorem

We shall use the following result of [12]. Let $(\Omega,\ \Sigma,\ \mu_i)$ $(i = 1,\ \ldots,\ n)$ be probability spaces, and let Ω^n be the product space with product measure $P = \mu_1 \times \cdots \times \mu_n$. For $A \subseteq \Omega^n$ and $\omega = (\omega_1,\ \ldots,\ \omega_n) \in \Omega^n$. We denote $(\beta_1,\ \ldots,\ \beta_n)$ by β. Talagrand's convex distance is defined by

$$d_T(\omega,\ A) = \sup\left\{z(\beta)\colon (\beta_1,\ \ldots,\ \beta_n) \in \mathrm{R}^n,\ \sum_{i=1}^n \beta_i^2 = 1\right\}$$

where

$$z(\beta) = \inf\left\{\sum_{i=1}^{n} \beta_i \cdot I(\omega_i \neq a_i)\colon (a_1, \ldots, a_n) \in A\right\}$$

and $I(\omega_i \neq a_i) = 1$ if $\omega_i \neq a_i$; zero, otherwise.

Theorem 2.3 (Talagrand, [12, Theorem 4.1.1]): For any $A \subseteq \Omega^n$ with $P(A) > 0$

$$\int_{\Omega^n} \exp\left(\frac{1}{4}\, d_T(\omega, A)^2\right) dP(\omega) \leq \frac{1}{P(A)}.$$

Consequently

$$P\left(d_T(\omega, A) \geq t\right) \leq \frac{1}{P(A)} \cdot \exp\left(-\frac{t^2}{4}\right). \qquad \square$$

III. Deviation Bound for the Maximum Squared Error $Q_{\max}(n)$

Recall that the input for wavelet shrinkage is $(y_1, \ldots, y_n)$, where $y_i = f_i + e_i$ $(i = 1, \ldots, n)$, the f_i are samples from the original signal f, and the e_i are additive noise. The e_i are independent random variables. We assume that the noise is bounded (with $|e_i| \leq \frac{b}{2}$), so each random variable e_i is a measurable function

$$e_i\colon \omega_i \in \Omega \mapsto e_i(\omega_i) \in \left[-\frac{b}{2}, \frac{b}{2}\right].$$

Accordingly, we view $(e_1, \ldots, e_n)$ as a function

$$\omega = (\omega_1, \ldots, \omega_n) \in \Omega^n \mapsto e(\omega) = (e_1(\omega_1), \ldots, e_n(\omega_n)) \in \left[-\frac{b}{2}, \frac{b}{2}\right]^n.$$

To simplify the notation we often write $e_i(\omega)$ for $e_i(\omega_i)$.

In what follows, the proofs are technical. But there is a clear guideline. To apply Talagrand's theorem we need a subset A of Ω^n which has positive probability. Moreover, the maximum squared error for noise $e(\omega)$, $\omega \in A$, is small. A technical difficulty is how to construct such a subset A. We do this in Section III-A. After A is obtained, in Section III-B, we expand A by Talagrand's convex distance. We achieve this by relating the discrete wavelet transform to Talagrand's convex distance.

A. The Subset A

Recall that we assume $n = 2^J$. For any $\omega \in \Omega^n$ we decompose the noise sample sequence $e(\omega)$ into blocks of length J, as follows:

$$e(\omega) = \left(\ldots, e_{kJ+1}(\omega), \ldots, e_{(k+1)J}(\omega), \ldots\right)$$

where $k = 0, \ldots, \frac{1}{J}\, 2^J - 1$. Here, for simplicity, we regard $\frac{1}{J}\, 2^J = 2^{J - \log J}$ as an integer (i.e., we assume that J is a power of 2).

For the Haar wavelets we define the subset $A \subset \Omega^n$ as follows:

$$A = \left\{\omega \in \Omega^n\colon (\forall \ell \in [-1, J - \log J]) \cdot \left(\forall k \in \left[0, 2^{J - \log J - \ell} - 1\right]\right), \cdot \left|\sum_{i=0}^{J2^{\ell-1}-1} e_{k2^\ell J + i + 1}(\omega)\right| \leq bJ2^{\ell/2}\sqrt{2^{-1}\ln 2}\right\}.$$

For the interval wavelet system we define

$$A = \left\{\omega \in \Omega^n\colon (\forall \ell \in [-1, J - \log J]) \cdot \left(\forall k \in \left[0, 2^{J - \log J - \ell} - 1\right]\right), \left|\sum_{i=0}^{J2^{\ell}-1} e_{k2^\ell J + i + 1}(\omega) \cdot \alpha_{i,\, J - \log J - \ell,\, k}\right| \leq bJ2^{\ell/2}\sqrt{2^{-1}\ln 2} \quad \text{and} \quad \left|\sum_{i=0}^{J2^{\ell}-1} e_{k2^\ell J + i + 1}(\omega) \cdot \beta_{i,\, J - \log J - \ell,\, k}\right| \leq bJ2^{\ell/2}\sqrt{2^{-1}\ln 2}\right\}.$$

The following lemma shows that A has positive probability measure, a desired property. To prove this lemma we need a classical result from probability theory.

Theorem 3.1 (Hoeffding's inequality): Let $X_1, \ldots, X_m$ be independent random variables with $b_1 \leq X_i \leq b_2$ $(i = 1, \ldots, m)$. Then for all $t > 0$

$$P\left(\left|\sum_{i=1}^{m} (X_i - E[X_i])\right| \geq t\right) \leq \exp\left(-\frac{2t^2}{m(b_2 - b_1)^2}\right). \qquad \square$$

Lemma 3.2: For all $n > 1$, $P(A) \geq 1 - \frac{4}{\log n} + \frac{1}{n}$ for the Haar wavelets, and $P(A) \geq 1 - \frac{8}{\log n} + \frac{2}{n}$ for the interval wavelet system.

In either case, if $n \geq 256$ then $P(A) \geq \frac{1}{128}$. If $n \geq 2^9$ then $P(A) > \frac{1}{9}$. Moreover, $P(A)$ tends to 1 when $n \to \infty$.

Proof: We first give the proof for the Haar wavelets. For any $\ell \in [-1, J - \log J]$ and $k \in [0, 2^{J - \log J - \ell} - 1]$, the noise samples $e_{k2^\ell J + 1}, \ldots, e_{(k+1)2^\ell J}$ are independent random variables, each with values in $[-\frac{b}{2}, \frac{b}{2}]$. So Hoeffding's inequality applies, and since $E[e_i] = 0$ for all i, we obtain for all $t > 0$

$$P\left(\left|\sum_{i=0}^{2^{\ell-1}J-1} e_{k2^\ell J + i + 1}\right| \leq t\right) \geq 1 - \exp\left(-\frac{2t^2}{2^\ell J b^2}\right).$$

Letting $t = b2^{\ell/2}J\sqrt{2^{-1}\ln 2}$ we obtain

$$P\left(\left|\sum_{i=0}^{2^{\ell-1}J-1} e_{k2^\ell J + i + 1}\right| \leq b2^{\ell/2}J\sqrt{2^{-1}\ln 2}\right) \geq 1 - \frac{1}{n}. \qquad (1)$$

For $\ell \in [-1, J - \log J]$ and $k \in [0, 2^{J - \log J - \ell} - 1]$, let

$$A_{\ell,\,k} = \left\{\omega \in \Omega^n\colon \left|\sum_{i=0}^{2^{\ell-1}J-1} e_{k2^\ell J + i + 1}(\omega)\right| \leq b2^{\ell/2}J\sqrt{2^{-1}\ln 2}\right\}$$

and let $A_\ell = \bigcap_{k=0}^{2^{J - \log J - \ell} - 1} A_{\ell,\,k}$.

Then by (1), $P(A_{\ell,\,k}) \geq 1 - \frac{1}{n}$. For the complements of these sets we have

$$\overline{A}_\ell = \bigcup_{k=0}^{2^{J - \log J - \ell} - 1} \overline{A}_{\ell,\,k},$$

and hence,

$$P(\overline{A}_\ell) \leq \sum_{k=0}^{2^{J - \log J - \ell} - 1} \frac{1}{n}.$$

Since $n = 2^J$ we obtain

$$P(\overline{A}_\ell) \leq \frac{2^{-\ell}}{\log n}.$$

Since $A = \bigcap_{\ell=-1}^{J - \log J} A_\ell$, we have

$$P(A) \geq 1 - \sum_{\ell=-1}^{J - \log J} P(\overline{A}_\ell) \geq 1 - \sum_{\ell=-1}^{J - \log J} \frac{2^{-\ell}}{\log n}.$$

Hence, $P(A) \geq 1 - \frac{4}{\log n} + \frac{1}{n}$. This proves the Lemma for the Haar case.

For the interval wavelet system we let

$$A^\alpha = \Bigg\{ \omega \in \Omega^n : \\ (\forall\, \ell \in [-1,\, J - \log J]) \left(\forall\, k \in \left[0,\, 2^{J-\log J-\ell} - 1\right]\right) \\ \cdot \left| \sum_{i=0}^{J2^\ell - 1} e_{k2^\ell J+i+1}(\omega) \cdot \alpha_{i,\, J-\log J-\ell,\, k} \right| \\ \le\, bJ2^{\ell/2}\sqrt{2^{-1}\ln 2} \Bigg\}$$

and

$$A^\beta = \Bigg\{ \omega \in \Omega^n : \\ (\forall\, \ell \in [-1,\, J - \log J]) \left(\forall\, k \in \left[0,\, 2^{J-\log J-\ell} - 1\right]\right) \\ \cdot \left| \sum_{i=0}^{J2^\ell - 1} e_{k2^\ell J+i+1}(\omega) \cdot \beta_{i,\, J-\log J-\ell,\, k} \right| \\ \le\, bJ2^{\ell/2}\sqrt{2^{-1}\ln 2} \Bigg\}.$$

Then $A = A^\alpha \cap A^\beta$.

We also let

$$A^\alpha_{\ell,\,k} = \left\{ \omega \in \Omega^n : \left| \sum_{i=0}^{J2^\ell - 1} e_{k2^\ell J+i+1}(\omega) \cdot \alpha_{i,\, J-\log J-\ell,\, k} \right| \le bJ2^{\ell/2}\sqrt{2^{-1}\ln 2} \right\}$$

and

$$A^\beta_{\ell,\,k} = \left\{ \omega \in \Omega^n : \left| \sum_{i=0}^{J2^\ell - 1} e_{k2^\ell J+i+1}(\omega) \cdot \beta_{i,\, J-\log J-\ell,\, k} \right| \le bJ2^{\ell/2}\sqrt{2^{-1}\ln 2} \right\}.$$

Moreover, we let

$$A^\alpha_\ell = \bigcap_k A^\alpha_{\ell,\,k} \quad \text{and} \quad A^\beta_\ell = \bigcap_k A^\beta_{\ell,\,k}.$$

Then $A_\ell = A^\alpha_\ell \cap A^\beta_\ell$, hence, $\overline{A}_\ell = \overline{A}^{\,\alpha}_\ell \cup \overline{A}^{\,\beta}_\ell$.

By the same proof as for Haar wavelets above: $P(\overline{A}^{\,\alpha}_\ell)$ and $P(\overline{A}^{\,\beta}_\ell) \le \frac{2^{-\ell}}{\log n}$. Hence,

$$P(\overline{A}_\ell) \le \frac{2^{-\ell+1}}{\log n}.$$

Since $A = \bigcap_{\ell=-1}^{J-\log J} A_\ell$, we obtain by a similar calculation as in the Haar case

$$P(A) \ge 1 - \frac{8}{\log n} + \frac{2}{n}. \qquad \square$$

Lemma 3.3: For all $\omega \in A$, all $j \in]J_0,\, J[$, and all $k \in [0,\, 2^j - 1]$, we have (for some constant $C_\varphi \ge 1$, depending only on the wavelet system)

$$\left| d^{(e(\omega))}_{j,\,k} \right| \le bC_\varphi \sqrt{\frac{\log n}{n}}$$

and for all $k \in [0,\, 2^{J_0} - 1]$

$$\left| c^{(e(\omega))}_{J_0,\,k} \right| \le bC_\varphi \sqrt{\frac{\log n}{n}}\,.$$

Proof: We consider two cases for j.

Case 1.: $J_0 \le j \le J - \log J + 1$. We write j as $J - \log J - \ell$, where $-1 \le \ell \le J - \log J - J_0$. Let us first consider Haar wavelets. By (H2) (in Proposition 2.1) we have

$$d^{(e(\omega))}_{j,\,k} = 2^{-J+j/2} \left(\sum_{i=0}^{2^{\ell-1}J-1} e_{k2^\ell J+i+1}(\omega) - \sum_{i=0}^{2^{\ell-1}J-1} e_{(k+1/2)2^\ell J+i+1}(\omega) \right).$$

Since $\omega \in A$ we can apply the defining property of A to

$$\left| \sum_{i=0}^{J2^{\ell-1}-1} e_{i+1+k2^\ell J} \right| = \left| \sum_{i=0}^{J2^{\ell-1}-1} e_{i+1+2k2^{\ell-1}J} \right|.$$

Since $2k$ is in the correct range $[0,\, 2^{j+1} - 2] = [0,\, \frac{1}{J}\, 2^{J-(\ell-1)} - 2]$, we have

$$\left| \sum_{i=0}^{J2^{\ell-1}-1} e_{i+1+k2^\ell J} \right| \le bJ2^{(\ell-1)/2}\sqrt{2^{-1}\ln 2}.$$

Similarly

$$\left| \sum_{i=0}^{J2^{\ell-1}-1} e_{i+1+(k+\frac{1}{2})2^\ell J} \right| = \left| \sum_{i=0}^{J2^{\ell-1}-1} e_{i+1+(2k+1)2^{\ell-1}J} \right| \le bJ2^{(\ell-1)/2}\sqrt{2^{-1}\ln 2}$$

we used the defining property of A, since the range of $2k+1$ is

$$[0,\, 2^{j+1} - 2 + 1] = [0,\, \tfrac{1}{J}\, 2^{J-(\ell-1)} - 1].$$

By combining these two bounds we obtain

$$\left| d^{(e(\omega))}_{j,\,k} \right| \le 2^{-J+j/2} \cdot 2 \cdot bJ2^{(\ell-1)/2}\sqrt{2^{-1}\ln 2} < b\sqrt{\ln 2}\sqrt{\frac{\log n}{n}} \le b\sqrt{\frac{\log n}{n}}\,.$$

Let us now consider Case 1 for the interval wavelet system. By (D2) in (Proposition 2.2)

$$d^{(e(\omega))}_{j,\,k} = 2^{-J+j/2} \cdot \sum_{i=0}^{2^\ell J-1} e_{k2^\ell J+i+1}(\omega) \cdot \beta_{i,\,j,\,k}\,.$$

Since $\omega \in A$

$$\left| d^{(e(\omega))}_{j,\,k} \right| \le 2^{-J+j/2} \cdot bJ2^{(\ell-1)/2}\sqrt{2^{-1}\ln 2} = b2^{(-J+\log J)/2}\sqrt{2^{-1}\ln 2} = b\sqrt{\frac{\log n}{n}}\,\sqrt{2^{-1}\ln 2} \le b\sqrt{\frac{\log n}{n}}\,.$$

Case 2 : $J - \log J + 2 \le j < J$. For the Haar wavelets we use the boundedness of the noise $|e_i - e_j| \le b$. Hence, by (H2)

$$\left|d_{j,k}^{(e(\omega))}\right| \le 2^{-J+j/2} b(J2^{\ell-1} - 1) \le b\sqrt{\frac{\log n}{n}}.$$

For the interval wavelet system, (D2) yields

$$\begin{aligned}\left|d_{j,k}^{(e(\omega))}\right| &\le 2^{-J+j/2} \sum_{i=0}^{2^{J-j}-1} \left|e_{k2^{\ell}J+i+1}(\omega)\right| \cdot |\beta_{i,j,k}| \\ &= 2^{-J+j/2}\, 2^{J-j}\, \frac{b}{2}\, C_\varphi \\ &\le \frac{b}{2} C_\varphi\, 2^{-j/2} \le bC_\varphi \sqrt{\frac{\log n}{n}}\end{aligned}$$

by using $j \ge J - \log J + 2$ for the last inequality.

By an argument similar to the above we obtain the bound for $|c_{J_0,k}^{(e(\omega))}|$. □

To implement wavelet shrinkage we need two parameters: a decomposition level J_0 and a threshold $\lambda_{n,\delta}$. We define

$$J_1 = \left\lceil \frac{1}{1+2\alpha}(J - \log J) \right\rceil$$

and we choose J_0 so that $J_0 \le J_1$.

For the Haar wavelets (when $0 < \alpha \le 1$) we can simply pick $J_0 = 0$, but for the interval wavelet system (when $1 < \alpha$ and we have $N = \lceil \alpha \rceil$ vanishing moments), we also require (see [4]) that $J_0 \ge 1 + \log(2N-1)$. When $\alpha > 1$ we choose

$$J_0 = 1 + \lceil \log(2\lceil \alpha \rceil - 1) \rceil.$$

Thus, for J_0 to exist (when $\alpha > 1$) we need $n = 2^J$ to be such that

$$1 + \log(2\lceil \alpha \rceil - 1) \le J_1.$$

A sufficient condition for this is that

$$J - \log J \ge (1 + \log(2\alpha + 1))(1 + 2\alpha)$$

or equivalently

$$\frac{n}{\log n} \ge (4\alpha + 2)^{2\alpha+1}.$$

By using the fact that $\frac{n}{\log n}$ is an increasing function of n and that the relation $\frac{y}{\log y} \ge x$ is implied by $y \ge x \cdot \log x \cdot \log\log x$, we have the following sufficient condition on n.

When $\alpha > 1$ we assume that

$$n \ge (4\alpha + 2)^{2\alpha+2} \cdot (\log(4\alpha + 2))^2.$$

We use the threshold

$$\lambda_{n,\delta} = C_\varphi b \left(1 + 2\sqrt{(1+\delta)\ln 2}\right) \sqrt{\frac{\log n}{n}}.$$

The first step of the wavelet shrinkage algorithm is DWT, which maps $(y_1, \ldots, y_n)$ to

$$\begin{aligned}\sqrt{n}\,(&c_{J_0,0}^{(y)}, \ldots, c_{J_0,2^{J_0}-1}^{(y)}, d_{J_0,0}^{(y)}, \ldots, \\ &d_{J_0,2^{J_0}-1}^{(y)}, \ldots, d_{J-1,0}^{(y)}, \ldots, d_{J-1,2^{J-1}-1}^{(y)})\end{aligned}$$

where $n = 2^J$.

Since $y_i = f_i + e_i$ and the DWT is linear we have

$$c_{J_0,k}^{(y)} = c_{J_0,k}^{(f)} + c_{J_0,k}^{(e)}, \qquad 0 \le k < 2^{J_0}$$

and

$$d_{j,k}^{(y)} = d_{j,k}^{(f)} + d_{j,k}^{(e)}, \qquad J_0 \le j < J,\ 0 \le k < 2^j$$

where $c_{J_0,k}^{(f)}$, $d_{j,k}^{(f)}$ and $c_{J_0,k}^{(e)}$, $d_{j,k}^{(e)}$ are the wavelet coefficients for $(f_1, \ldots, f_n)$ and $(e_1, \ldots, e_n)$, respectively.

The second step of wavelet shrinkage is thresholding. We shall prove our result for soft thresholding. But in our proofs it will be easy to see that our results will hold for hard thresholding as well. For soft thresholding, we have

$$\tilde{d}_{j,k} = \begin{cases} d_{j,k}^{(y)} - \lambda_{n,\delta}, & \text{if } d_{j,k}^{(y)} > \lambda_{n,\delta} \\ 0, & \text{if } |d_{j,k}^{(y)}| \le \lambda_{n,\delta} \\ d_{j,k}^{(y)} + \lambda_{n,\delta}, & \text{if } d_{j,k}^{(y)} < -\lambda_{n,\delta}. \end{cases}$$

The last step of wavelet shrinkage is the inverse of DWT which yields $\tilde{y} = (\tilde{y}_1, \ldots, \tilde{y}_n)$. If we let

$$\tilde{y}(x) = \sum_{k=0}^{2^{J_0}-1} c_{J_0,k}^{(y)} \varphi_{J_0,k}(x) + \sum_{j=J_0}^{J-1} \sum_{k=0}^{2^j-1} \tilde{d}_{j,k} \psi_{j,k}(x) \qquad (2)$$

then we obtain $\tilde{y}_i = \tilde{y}(\frac{i}{n})$ for $i = 1, \ldots, n$.

B. Expanding the Subset A

Let W be the orthogonal matrix that represents the DWT. Let $A \subseteq \Omega^n$ be as above. For any $\delta > 0$, we define the following subset of Ω^n:

$$B_\delta = \left\{ \omega' \in \Omega^n \colon (\forall\, \ell \in [1, n]),\ \inf_{\omega \in A} \left| \sum_{i=1}^{n} W_{\ell,i} \left(e_i(\omega') - e_i(\omega)\right) \right| < 2b\sqrt{(1+\delta)\ln n} \right\}.$$

We shall show that B_δ contains a subset that is an expansion of the subset A within Talagrand's convex distance, and thus has probability measure quite close to one (Lemma 3.5). Furthermore, for every ω' in B_δ, the wavelet shrinkage works as well as it works for ω in A (Lemma 3.6).

Lemma 3.4: For all $\omega' \in B_\delta$ and all $k \in [0, 2^{J_0} - 1]$: $|c_{J_0,k}^{(e(\omega'))}| \le \lambda_{n,\delta}$.

For all $j \in [J_0, J-1]$ and $k \in [0, 2^j - 1]$: $|d_{j,k}^{(e(\omega'))}| \le \lambda_{n,\delta}$.

Proof: By the definition of B_δ, for every $\omega' \in B_\delta$ there exists $\omega \in A$ such that

$$\sqrt{n}\left|c_{J_0,k}^{(e(\omega))} - c_{J_0,k}^{(e(\omega'))}\right| \le b2\sqrt{(1+\delta)\ln n}$$

and

$$\sqrt{n}\left|d_{j,k}^{(e(\omega))} - d_{j,k}^{(e(\omega'))}\right| \le b2\sqrt{(1+\delta)\ln n}.$$

The lemma then follows from Lemma 3.3. □

For the following lemma we use the threshold $\lambda_{n,\delta}$ as above; we let $n_0 = 2^9$ when $0 < \alpha \le 1$, and $n_0 = (4\alpha+2)^{2\alpha+2} \cdot (\log(4\alpha+2))^2$ when $\alpha > 1$.

Lemma 3.5: When $n \ge n_0$, $P(B_\delta) > 1 - \frac{9}{n^{1+\delta}}$.

Proof: We first prove that

$$\left\{\omega' \in \Omega^n \colon d_T(\omega', A) < 2\sqrt{(1+\delta)\ln n}\right\} \subseteq B_\delta.$$

Recall the definition

$$d_T(\omega', A) = \sup\left\{ z(\beta) \colon (\beta_1, \ldots, \beta_n) \in \mathrm{R}^n,\ \sum_{i=1}^{n} \beta_i^2 = 1 \right\}$$

where

$$z(\beta) = \inf\left\{ \sum_{i=1}^{n} \beta_i \cdot I(\omega_i' \ne a_i) \colon (a_1, \ldots, a_n) \in A \right\}.$$

We choose the following n vectors for $\beta = (\beta_1, \ldots, \beta_n)$ in the above formula:

$$(|W_{1,\ell}|, \ldots, |W_{n,\ell}|), \qquad \text{for } \ell = 1, \ldots, n.$$

Since W is orthogonal, all its row vectors have unit length. For all $\omega' \in \Omega^n$, $\omega = (\omega_1, \ldots, \omega_n) \in A$, and $1 \le \ell \le n$ we have

$$\left|\sum_{i=1}^{n} W_{i,\ell}\left(e_i(\omega') - e_i(\omega)\right)\right| \le b\sum_{i=1}^{n} |W_{i,\ell}| \cdot I\left(e_i(\omega') \ne e_i(\omega)\right) \le b\sum_{i=1}^{n} |W_{i,\ell}| \cdot I(\omega' \ne \omega).$$

(The last inequality follows from the fact that $I(e_i(\omega') \ne e_i(\omega)) \le I(\omega' \ne \omega)$, because $e_i(\omega') \ne e_i(\omega)$ implies $\omega' \ne \omega$.)

Hence, for all $\omega' \in \Omega^n$ and $1 \le \ell \le n$

$$\inf\left\{\left|\sum_{i=1}^{n} W_{i,\ell}\left(e_i(\omega') - e_i(\omega)\right)\right| : \omega \in A\right\} \le \inf\left\{\sum_{i=1}^{n} |W_{i,\ell}| \cdot I(\omega' \ne \omega) b : \omega \in A\right\} = b\inf\left\{\sum_{i=1}^{n} |W_{i,\ell}| \cdot I(\omega' \ne \omega) : \omega \in A\right\}.$$

Therefore, if $d_T(\omega', A) \le 2\sqrt{(1+\delta)\ln n}$ then for all $1 \le \ell \le n$

$$\inf\left\{\left|\sum_{i=1}^{n} W_{i,\ell}\left(e_i(\omega') - e_i(\omega)\right)\right| : \omega \in A\right\} \le b2\sqrt{(1+\delta)\ln n}.$$

This means that $\omega' \in B_\delta$, and this proves that

$$\left\{\omega' \in \Omega^n : d_T(\omega', A) < 2\sqrt{(1+\delta)\ln n}\right\} \subseteq B_\delta.$$

Hence,

$$P(B_\delta) \ge P\left(\left\{\omega' \in \Omega^n :; d_T(\omega', A) < 2\sqrt{(1+\delta)\ln n}\right\}\right).$$

By Talagrand's theorem this means

$$P(B_\delta) \ge 1 - \frac{1}{P(A)} \cdot \exp\left(-(1+\delta)\ln 2\right) > 1 - \frac{9}{n^{1+\delta}}. \quad \square$$

Lemma 3.6: For all $\omega' \in B_\delta$ we have

i) When $J_1 \le j < J$, $0 \le k < 2^j$

$$\left|\tilde{d}_{j,k}(\omega') - d^{(f)}_{j,k}\right| \le \left|d^{(f)}_{j,k}\right| \le C_\varphi M \cdot 2^{-j\left(\frac{1}{2}+\alpha\right)}.$$

ii) When $J_0 \le j < J_1$, $0 \le k < 2^j$

$$\left|\tilde{d}_{j,k}(\omega') - d^{(f)}_{j,k}\right| \le 2\lambda_{n,\delta}.$$

Proof: To prove i), we note first that by (H3), (D3) we have $|d^{(f)}_{j,k}| \le C_\varphi M\ 2^{-j(1/2+\alpha)}$.

To prove the inequality $|d^{(f)}_{j,k} - \tilde{d}_{j,k}| \le |d^{(f)}_{j,k}|$ one considers six cases, according to the possible relative positions of 0, $d^{(f)}_{j,k}$, and $\tilde{d}_{j,k}$. If $0 \le \tilde{d}_{j,k} \le d^{(f)}_{j,k}$, or if $d^{(f)}_{j,k} \le \tilde{d}_{j,k} \le 0$, the inequality is obvious from the order picture. The other four cases are not possible, since they would imply that $|d^{(e(\omega))}_{j,k}| > \lambda_{n,\delta}$, contradicting what we saw a little earlier. This proves i).

For the proof of ii) we consider two cases. If $\tilde{d}_{j,k} = 0$, $|d^{(y)}_{j,k}| \le \lambda_{n,\delta}$, hence,

$$|d^{(f)}_{j,k} - \tilde{d}_{j,k}| = |d^{(f)}_{j,k}| = |d^{(y)}_{j,k} - d^{(e)}_{j,k}| \le |d^{(y)}_{j,k}| + |d^{(e)}_{j,k}| \le \lambda_{n,\delta} + \lambda_{n,\delta}.$$

In the second case, $|d^{(y)}_{j,k}| > \lambda_{n,\delta}$, and

$$|d^{(f)}_{j,k} - \tilde{d}_{j,k}| = |d^{(e)}_{j,k} - \lambda_{n,\delta}| \le \lambda_{n,\delta} + \lambda_{n,\delta}.$$

This proves ii). $\square$

We are ready to give *a proof of* Theorem 1.1 as follows.

Proof: At the beginning of Section II-A we defined the function $\overline{f}$, and its wavelet coefficients. We have

$$\overline{f}(x) = \sum_{k=0}^{2^{J_0}-1} c^{(f)}_{J_0,k}\varphi_{J_0,k}(x) + \sum_{j=J_0}^{J_1-1}\sum_{k=0}^{2^j-1} d^{(f)}_{j,k}\psi_{j,k}(x) + \sum_{j=J_1}^{J-1}\sum_{k=0}^{2^j-1} d^{(f)}_{j,k}\psi_{j,k}(x),$$

and $f_i = \overline{f}\left(\frac{i}{n}\right)$ for $1 \le i \le n$.

In connection with the thresholding of y we define the function

$$\tilde{y}(x) = \sum_{k=0}^{2^{J_0}-1} c^{(y)}_{J_0,k}\varphi_{J_0,k}(x) + \sum_{j=J_0}^{J_1-1}\sum_{k=0}^{2^j-1} \tilde{d}_{j,k}\psi_{j,k}(x) + \sum_{j=J_1}^{J-1}\sum_{k=0}^{2^j-1} \tilde{d}_{j,k}\psi_{j,k}(x).$$

By Lemma 3.4, we have for all $\omega' \in B_\delta$

$$\left|c^{(y)}_{J_0,k} - c^{(f)}_{J_0,k}\right| = \left|c^{(e(\omega'))}_{J_0,k}\right| \le \lambda_{n,\delta}. \tag{3}$$

By Lemma 3.6, we have for all $\omega' \in B_\delta$, for $J_1 \le j < J$, $0 \le k < 2^j$

$$\left|\tilde{d}_{j,k} - d^{(f)}_{j,k}\right| \le \left|d^{(f)}_{j,k}\right| \le C_\varphi M \cdot 2^{-j\left(\frac{1}{2}+\alpha\right)} \tag{4}$$

and for $J_0 \le j < J_1$, $0 \le k < 2^j$

$$\left|\tilde{d}_{j,k} - d^{(f)}_{j,k}\right| \le 2\lambda_{n,\delta}. \tag{5}$$

Let us first deal with the case of Haar wavelets (when $\alpha \le 1$). For a given j, the supports of different Haar wavelets do not overlap. Therefore, for all $x \in]0, 1]$ there exist K_1 and $K(j)$ such that

$$\left|\tilde{f}(x) - \tilde{y}(x)\right| \le \left|c^{(y)}_{J_0,K_1} - c^{(f)}_{J_0,K_1}\right| \cdot 2^{J_0/2} + \sum_{j=J_0}^{J_1-1}\left|\tilde{d}_{j,K(j)} - d^{(f)}_{j,K(j)}\right| \cdot 2^{j/2} + \sum_{j=J_1}^{J-1}\left|\tilde{d}_{j,K(j)} - d^{(f)}_{j,K(j)}\right| \cdot 2^{j/2}.$$

This and (3)–(5) imply for all $x \in]0, 1]$

$$\left|\tilde{f}(x) - \tilde{y}(x)\right| \le (C_1 + C_2 + C_3) \cdot \left(\frac{\log n}{n}\right)^{\frac{\alpha}{1+2\alpha}} = \left(c_1' + c_2'\sqrt{1+\delta}\right) \cdot \left(\frac{\log n}{n}\right)^{\frac{\alpha}{1+2\alpha}}.$$

Letting $x = \frac{i}{n}$ $(1 \le i \le n)$ we obtain for all $\omega' \in B_\delta$

$$\left|f_i - \tilde{y}_i(\omega')\right| = \left|\tilde{f}\left(\frac{i}{n}\right) - \tilde{y}\left(\frac{i}{n}\right)\right| \le \left(c_1' + c_2'\sqrt{1+\delta}\right) \cdot \left(\frac{\log n}{n}\right)^{\frac{\alpha}{1+2\alpha}}.$$

In the Haar case, the theorem follows from this and the fact that $P(B_\delta) > 1 - \frac{9}{n^{1+\delta}}$ (when $n \ge n_0$).

For wavelets on the interval (when $\alpha > 1$ and the number of vanishing moments is $N = \lceil \alpha \rceil$), there are never more than $2N$ wavelets that overlap (for a given j). Indeed, in the above sums we have for each j and each x: $0 \le 2^j x - k \le 2N - 1$. (Other values of k would place the argument $2^j x - k$ of the wavelet functions outside of the support and would hence only produce zero-terms in the sums.) Hence, k only needs to range from $\lceil 2^j x \rceil - 2N + 1$ through $\lceil 2^j x \rceil$, which corresponds to $2N$ values of k.

Therefore, the same calculation as for Haar wavelets applies, except that the constants C_1, C_2, C_3, c_1', and c_2' need to be multiplied by $2N$.□

ACKNOWLEDGMENT

D. Hong wishes to thank Ronald DeVore for his advice and Shushuang Man for helpful discussions while writing this correspondence.

REFERENCES

[1] R. Averkamp and Ch. Houdré, "Wavelet thresholding for non (necessarily) Gaussian noise: Idealism," preprint, [Online]. Available: http://www.math.gatech.edu/houdre/.

[2] ——, "Wavelet thresholding for non (necessarily) Gaussian noise: Functionality," preprint, [Online]. Available: http://www.math.gatech.edu/houdre/.

[3] L. D. Brown and M. G. Low, "Superefficiency and lack of adaptability in functional estimation," manuscript.

[4] A. Cohen, I. Daubechies, B. Jawerth, and P. Vial, "Multiresolution analysis, wavelets and fast algorithms on an interval," *C. R. l'Acad. Sci. de Paris*, ser. I, vol. 316, pp. 417–421, 1993.

[5] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA: SIAM, 1992.

[6] D. Donoho and I. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[7] D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard, "Wavelet shrinkage: Asymptopia?," *J. Roy. Statist. Soc.*, ser. B, vol. 57, no. 2, pp. 301–369, 1995.

[8] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, vol. 58, pp. 13–30, 1965.

[9] O. V. Lepskii, "On one problem of adaptive estimation on white Gaussian noise" (in Russian), *Teor. Veoryatnost. i Premenen.*, vol. 35, pp. 459–470, 1990. English translation: *Theory Probab. Applic.*, vol. 35, pp. 454–466, 1990.

[10] Y. Meyer, *Wavelets and Operators*. Cambridge, U.K.: Cambridge Univ. Press, 1992.

[11] J. M. Steele, *Probability Theory and Combinatorial Optimization*. Philadelphia, PA: SIAM, 1997.

[12] M. Talagrand, "Concentration of measure and isoperimetric inequalities in product spaces," *Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques*, vol. 81, pp. 73–205, 1995.

[13] B. Vidakovic, *Statistical Modeling by Wavelets*. New York: Wiley, 1999.

A New Metric for Probability Distributions

Dominik M. Endres and Johannes E. Schindelin

***Abstract*—We introduce a metric for probability distributions, which is bounded, information-theoretically motivated, and has a natural Bayesian interpretation. The square root of the well-known χ^2 distance is an asymptotic approximation to it. Moreover, it is a close relative of the capacitory discrimination and Jensen–Shannon divergence.**

***Index Terms*—Capacitory discrimination, χ^2 distance, Jensen–Shannon divergence, metric, triangle inequality.**

I. INTRODUCTION

This correspondence is the result of the authors' search for a probability metric that is bounded and can be easily interpreted in terms of both information-theoretical and probabilistic concepts. Metric properties are the prerequisites for several important convergence theorems for iterative algorithms, i.e., Banach's fixed point theorem [2], which is the basis of several pattern-matching algorithms. Boundedness is a valuable property, too, when numerical applications are considered.

We will limit the following discussion to discrete probability distributions, but the result can be generalized to probability density functions.

II. MOTIVATION

The motivation we are presenting in this section is aimed at providing the reader with an idea of the meaning of the metric. As such, it is not to be understood as a derivation in a strict mathematical sense. However, we will observe mathematical rigor in the following section, which contains the actual proof of the metric properties.

Let X be a discrete random variable which can take on N different values $\in \Omega_N = \{\omega_1, \ldots, \omega_N\}$. We now draw an independent and identically distributed (i.i.d.) sample $\tilde{X}$, where each observation is drawn from one of two known distributions, P and Q. Each of those is used with equal probability. However, we do not know which one is used when. Now we wish to find the coding strategy that gives the shortest average code length for the representation of the data. In other words, we are looking for the most *efficient* distribution R.

Let us call this code κ. The code lengths are $\kappa_i = -\log r_i$, where $i \in \{1, \ldots, N\}$ and r_i is the probability of $X = \omega_i$ under R. Denoting the expectation of κ with respect to (w.r.t.) P by $\mathcal{E}(\kappa, P)$, the average code length $\langle \kappa \rangle$ is then $\frac{1}{2}\mathcal{E}(\kappa, P) + \frac{1}{2}\mathcal{E}(\kappa, Q)$. By the very definition of the entropy, the *minimum* $\langle \kappa \rangle$ is obtained by setting $R = \frac{1}{2}(P + Q)$, i.e., $\langle \kappa \rangle = H(R)$.

An ideal observer, i.e., one who knows which distribution is used to generate the individual data, could reach an even shorter average code length $\frac{1}{2}H(P) + \frac{1}{2}H(Q)$. Hence, the redundancy of κ is $H(R) - \frac{1}{2}H(P) - \frac{1}{2}H(Q)$. The distance measure we studied is twice that redundancy

$$\begin{aligned} D_{PQ}^2 &= 2H(R) - H(P) - H(Q) \\ &= D(P\|R) + D(Q\|R) \\ &= \sum_{i=1}^{N} \left(p_i \log \frac{2p_i}{p_i + q_i} + q_i \log \frac{2q_i}{p_i + q_i} \right). \end{aligned} \tag{1}$$

Manuscript received May 6, 2002; revised February 28, 2003.

D. M. Endres is with the School of Psychology, University of St Andrews, St Andrews KY16 9JU, U.K. (e-mail: dme2@st-andrews.ac.uk).

J. E. Schindelin is with the Institut für Genetik, Biozentrum, Universität Würzburg, 97074 Würzburg, Germany (e-mail: gene099@mail.uni-wuerzburg.de).

Communicated by G. Lugosi, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Digital Object Identifier 10.1109/TIT.2003.813506