AVERAGES OF BEST WAVELET BASIS ESTIMATES FOR DENOISING

S. E. FERRANDO AND L. A. KOLASA

ABSTRACT. Donoho and Johnstone introduced an adaptive algorithm that extends nonlinear thresholding denoising in a fixed orthonormal basis to a multiple basis setting. In their work a search for an optimal basis from a large collection of orthonormal bases—i.e., a *library*—is introduced. That technique gives the so-called best ortho-basis estimate. In this paper we study the situation when many such libraries are available. We propose an algorithm that exploits the availability of many best ortho-basis approximations. The algorithm uses a strengthening of the convexity of the L^2 norm to produce an estimate which is an average of best ortho-basis estimates. Conditions under which the proposed algorithm offers improvements and corresponding numerical examples are also described.

1. INTRODUCTION

Suppose we have noisy data

$$y_j = f_j + \sigma\, \xi_j \quad j = 1, \dots , N, \tag{1}$$

where $f = (f_j)$ is a given signal and $\xi = (\xi_j)$ are samples from i.i.d.random variables with $\mathcal{N}(0, 1)$ distributions. Thresholding schemes on a *fixed* orthonormal basis [6], [7] are well known estimators for the underlying signal. It is also known [6], [7] that the quality of such denoising is related to how well the signal is compressed in the given orthonormal basis. It is natural to wonder, therefore, how to find the "best" orthonormal basis with which to denoise a given signal. In this case we say that the basis is *adapted* to the signal. Adaptive approximations necessarily offer a better compression of a signal than expansions in terms of fixed orthonormal basis. However, in the presence of noise, estimations by thresholding may not be improved by an adaptive expansion because the flexibility of the search may result in a basis that correlates well with the noise. Reference [5], building on results from [2], proposes an algorithm for adaptive denoising, namely an orthonormal basis which satisfies certain optimality conditions for denoising is chosen from a *library* of such bases. This scheme will be called *Best Ortho-Basis denoising* and its estimate *best wavelet basis estimate.* Implicit in this scheme is the possibility of using many such libraries. If this is actually the case, a further use of adaptivity is to actually combine, in a meaningful way, some of the available estimates. In this paper we propose a new algorithm based on selectively averaging best wavelet bases

Date: May 15, 2000.

1991 *Mathematics Subject Classification.* 42C40, 93E11, 94A08.

Key words and phrases. Signal Denoising Wavelet Packets. L^2 spaces. Adaptive Basis Selection. Oracles for Adaptation. Thresholding of Wavelet Coefficients.

Research partially supported by NSERC.

estimates obtained from the best ortho-basis algorithm applied to each library from a given collection of libraries.

We now describe the motivations behind our approach. Donoho and Johnstone (D&J) select a basis $\hat{\mathcal{B}}_{\mathcal{L}}$ out of a library $\mathcal{L}$ by minimizing a certain entropy functional. In the case where many such libraries $\mathcal{L}_i$ and associated bases $\hat{\mathcal{B}}_{\mathcal{L}_i}$ are available a natural question arises: which of these bases gives rise to the best estimate? One possible approach is to combine these bases into a new library and apply D&J minimum entropy approach to this new collection. A better approach, which is the one taken in this paper, is to select a subset of the collection $\mathcal{T} \subseteq \{\mathcal{L}_i \in L\}$. It turns out that the elements from $\mathcal{T}$ offer the same advantages for denoising from the perspective of the D&J theory. We exploit this fact by taking an average of the estimates associated to $\mathcal{T}$. The fact that our average estimate can offer a better estimate than the ones associated to single elements of $\mathcal{T}$ is indicated by Proposition 1, which relies on the uniform convexity of the sphere in L^2 and D&J oracle inequalities.

The paper is organized as follows: Section 2 reviews the main results related to Best Ortho-Basis denoising; key notation is introduced along with critical remarks regarding a software implementation of this technique. Section 3 is the core of the paper; Proposition 1 is proved and then the *shell repelling algorithm*, based on this proposition, is described. Section 4 gives numerical examples, with discussions, related to the performance of the algorithms under a variety of conditions. Section 5 summarizes the main points and indicates possibilities for further research. Appendix A explores numerically the satisfability of the hypothesis needed to apply Proposition 1 for libraries of wavelet packets. Moreover, in this appendix we list the signals used in the numerical experiments.

2. Best Orthonormal Basis for Denoising

Here we summarize the main result from [5] as presented in [4]. Results for a single basis and for libraries will be presented simultaneously and contrasted. Suppose we have available a library $\mathcal{L}$ of orthogonal bases, such as the Wavelet Packet bases [2] or the Cosine Packet bases of Coifman and Meyer [1]. Let $\mathcal{B} \in \mathcal{L}$ and $\theta(x, \mathcal{B})$ denote the vector of coefficients of a vector x in the basis $\mathcal{B}$. Consider the family $\hat{\Phi}_{\mathcal{B}}$ of estimators defined by hard thresholding empirical coefficients in some basis $\mathcal{B} \in \mathcal{L}$. Such estimators $\hat{f}(y; z, \mathcal{B})$ in the coefficient domain are of the form

$$\theta_i(\hat{f}, \mathcal{B}) = z_i \; \theta_i(y, \mathcal{B}). \tag{2}$$

where each weight z_i is either 0 or 1. Formally, the set of estimators associated to $\mathcal{B}$ are $\hat{\Phi}_{\mathcal{B}} = \{\hat{f}(., z, \mathcal{B}) : z \in \{0,1\}^N\}$, and the ones associated to $\mathcal{L}$ are $\hat{\Phi}_{\mathcal{L}} = \bigcup_{\mathcal{B} \in \mathcal{L}} \hat{\Phi}_{\mathcal{B}} = \{\hat{f}(., z, \mathcal{B}) : z \in \{0,1\}^N, \mathcal{B} \in \mathcal{L}\}$. Given an estimator $\hat{f}$ of f, the quality of estimation is mesured in terms of its risk,

$$\mathcal{R}\left(\hat{f}(z, \mathcal{B}), f\right) = \mathbf{E}\left(||\hat{f}(y, z, \mathcal{B}) - f||^2\right). \tag{3}$$

Define the ideal risk for the case of a library by

$$\mathcal{R}_{\mathcal{L}}(f) = \inf_{\hat{f} \in \hat{\Phi}_{\mathcal{L}}} \mathcal{R}\left(\hat{f}(z, \mathcal{B}), f\right). \tag{4}$$

In the single basis setting the ideal risk is given by

$$\mathcal{R}_{\mathcal{B}}(f) = \inf_{\hat{f} \in \hat{\Phi}_{\mathcal{B}}} \mathcal{R}\left(\hat{f}(z,\mathcal{B}), f\right). \tag{5}$$

Notice that in order to attain these ideal risks, knowledge of the vector f is required. Estimates attaining these ideal risks are, therefore, not empirical estimates but *oracle* ones [5]. Associated to the ideal risk in (4) is the ideal basis $\mathcal{B}_{\mathcal{L}} \in \mathcal{L}$ defined by

$$\mathcal{B}_{\mathcal{L}} = \arg \inf_{\hat{f} \in \hat{\Phi}_{\mathcal{L}}} \mathcal{R}\left(\hat{f}(z,\mathcal{B}), f\right) = \arg \inf_{\hat{\Phi}_{\mathcal{B}} \subset \hat{\Phi}_{\mathcal{L}}} \mathcal{R}_{\mathcal{B}}(f) \tag{6}$$

In order to introduce empirical estimates, define, for $\lambda > 0$ given, the vector $\delta = \delta(y,\mathcal{B},\lambda) = (\delta_i(y,\mathcal{B},\lambda)$ by

$$\delta_i(y,\mathcal{B},\lambda) = 1_{\{|\theta(y,\mathcal{B})| > \sigma.\sqrt{\lambda}\}}. \tag{7}$$

Set $\lambda = \lambda_{\mathcal{B}} = 2\,\ln(N)$. The empirical estimate $\hat{f}(y;\delta(\mathcal{B},\lambda_{\mathcal{B}}))$ relative to $\mathcal{B}$ is given by:

$$\theta_i(\hat{f},\mathcal{B}) = \delta_i(y,\mathcal{B},\lambda_{\mathcal{B}}).\theta_i(y,\mathcal{B}). \tag{8}$$

The following oracle inequality holds for all f and $N > 4$ [6].

Theorem 1. *If* $\hat{f}_{\mathcal{B}} = \hat{f}_{\mathcal{B}}(y;\delta(\mathcal{B},\lambda_{\mathcal{B}}))$

$$\mathcal{R}(\hat{f}_{\mathcal{B}}, f) \leq 2\ln(N).\left(\sigma^2 + \mathcal{R}_{\mathcal{B}}(f)\right). \tag{9}$$

Equation (9) applied to $\mathcal{B} = \mathcal{B}_{\mathcal{L}}$ suggests that this ideal basis will deliver a better estimate than in any other basis $\mathcal{B} \in \mathcal{L}$. The main point of [5] is to give an algorithm to select a basis $\hat{\mathcal{B}}_{\mathcal{L}} \in \mathcal{L}$ such that behaves similarly to $\mathcal{B}_{\mathcal{L}}$ with respect to oracle inequalities. In order to describe these results, set $M_{\mathcal{L}}$ equal to the number of distinct vectors occuring among all bases in $\mathcal{L}$ and $t_{\mathcal{L}} = \sqrt{2\ln(M_{\mathcal{L}})}$. Choose $\xi > 8$ and set the threshold parameter to be $\lambda = \lambda_{\mathcal{L}} = (\xi.(1+t_{\mathcal{L}}))^2$. Define now the entropy functional

$$\mathcal{E}_{\lambda_{\mathcal{L}}}(y,\mathcal{B}) = \sum_{i=1}^{N} \min\left(\theta_i^2(y,\mathcal{B}), \sigma^2\,\lambda_{\mathcal{L}}\right). \tag{10}$$

Let $\hat{\mathcal{B}}_{\mathcal{L}}$ be the best (empirical) orthogonal basis relative to this entropy:

$$\hat{\mathcal{B}}_{\mathcal{L}} = \arg\min_{\mathcal{B} \in \mathcal{L}} \mathcal{E}_{\lambda_{\mathcal{L}}}(y,\mathcal{B}). \tag{11}$$

The risk of the empirical estimate satisfies the following oracle inequality for all f [4].

Theorem 2. *If* $\hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}}} = \hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}}}\left(y;\delta(\hat{\mathcal{B}}_{\mathcal{L}},\lambda_{\mathcal{L}})\right)$ *then*

$$\mathcal{R}(\hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}}}, f) \leq A(\xi).\lambda_{\mathcal{L}}.\left(\sigma^2 + \mathcal{R}_{\mathcal{L}}(f)\right), \tag{12}$$

where $A(\xi) = 6.(1 - 8/\xi)^{-1}$.

Remark 1. *In* [5] *a similar result is proved that holds with high probability, this is in contrast to the above result that holds in the mean.*

2.1. **Practical Considerations.** The parameters involved in Theorem 2 are not substantially larger than those appearing in Theorem 1. In practical situations though, the discrepancy in the values of λ may cause the best ortho-basis estimate to be poorer than single basis estimates. The following remarks are intended to remedy this problem. The threshold parameter $\lambda_{\mathcal{L}}$ plays a double role in the constructions described above. First, it is used to select $\hat{\mathcal{B}}$ and then it is used to threshold $\theta_i(y, \hat{\mathcal{B}})$. In this second application, $\lambda_{\mathcal{L}}$ is too large for the method to be competitive with thresholding in a single basis. In [5] it is mentioned that the parameter $\xi > 8$ could be made smaller. We have found that a good performance for the best wavelet basis denoising algorithm is obtained if one uses $\lambda = \lambda^\star = t_{\mathcal{L}}^2 = 2\ln(M_{\mathcal{L}})$ consistently in all computations. This choice has the appealing property that in the special case when the library consists of a single basis the parameter $\lambda^\star$ coincides with the $\lambda_{\mathcal{B}}$ used in Theorem 1.

3. Averages of Best Orthonormal Bases for Denoising

This section proposes an algorithm to compute a new estimate given as an average of previously found best empirical estimates. We will base our algorithm in Section 3.1 on Proposition 1. In some of the statements below we assume that the data $y = (y_j)$ in (1) is given and we will supress the dependency of $\mathcal{E}_{\lambda_{\mathcal{L}}}(y, \mathcal{B})$ on the data. We assume that a finite collection of libraries $L = \{\mathcal{L}_i\}$ is given; for simplicity, we will assume $M_{\mathcal{L}} = M_{\mathcal{L}'}$ if $\mathcal{L}, \mathcal{L}' \in L$ and set $M_L = M_{\mathcal{L}}$. All libraries used in Section 3 belong to L. We extend slightly the notation from Section 2, namely, $\xi > 8$ is fixed and set $t_L = t_{\mathcal{L}} = \sqrt{2\,\ln M_L}$ and $\lambda_L = \lambda_{\mathcal{L}} = (\xi(1+t_L))^2$.

Given, for each $\mathcal{L}_i \in L$, the best ortho-basis estimate $\hat{f}_i = \hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}_i}}$, we should like to reconstruct f as an average $\hat{f} = \frac{1}{p}\sum_{i=1}^p \hat{f}_i$ and maintain control over $\mathcal{R}\left(\hat{f}, f\right) = \mathbf{E}\left(||\frac{1}{p}\sum_{i=1}^p(\hat{f}_i - f)||^2\right)$ in terms of the $\mathcal{R}\left(\hat{f}_i, f\right)$. This is the purpose of Lemma 1.

Lemma 1. *Let x_i, $i = 1, \ldots, p$ be a collection of vectors in a real inner product space with squared norm $||x||^2 = \langle x, x\rangle$. Assume $r \le ||x_i|| \le R, \quad i = 1, \ldots, p$. Set $\epsilon_p^2 = \frac{1}{2}\sum_{i=1}^p\sum_{j=1}^p ||x_i - x_j||^2$; then*

$$(r^2 - \epsilon_p^2/p^2) \le ||\frac{1}{p}\sum_i^p x_i||^2 \le (R^2 - \epsilon_p^2/p^2). \tag{13}$$

Proof. This result is just a generalization of the parallelogram identity to the case of working with p vectors instead of 2.

$$p\sum_{i=1}^p ||x_i||^2 - ||\sum_{i=1}^p x_i||^2 = (p-1)\sum_{i=1}^p ||x_i||^2 - 2\sum_{i=1}^p\sum_{j=i+1}^p \langle x_i, x_j\rangle =$$

$$\sum_{i=1}^p\sum_{j=i+1}^p ||x_i - x_j||^2 = \frac{1}{2}\sum_{i=1}^p\sum_{j=1}^p ||x_i - x_j||^2 = \epsilon_p^2.$$

Clearly then the upper bound follows as,

$$||\frac{1}{p}\sum_{i=1}^p x_i||^2 = \frac{1}{p}\sum_{i=1}^p ||x_i||^2 - \epsilon_p^2/p^2 \le (R^2 - \epsilon_p^2/p^2).$$

The lower bound similarily follows. □

If we take $x_i = \hat{f}_i - f$ and inner product $\langle g, h \rangle = \mathbf{E}\left(\sum_{i=1}^{N} g_i h_i\right)$, then $||x_i||^2 = \mathcal{R}\left(\hat{f}_i, f\right)$. To have any meaningful control over $\mathcal{R}\left(\hat{f}, f\right) = ||\frac{1}{p}\sum_{i=1}^{p} x_i||^2$ then the constants r and R of Lemma 1 should not differ by too much; this says that each $\hat{f}_i$ should be in a thin annulus or *shell* centered at the original signal f. As it stands $R^2 = \max \mathcal{R}\left(\hat{f}_i, f\right)$ and $r^2 = \min \mathcal{R}\left(\hat{f}_i, f\right)$, so we ought to be more thoughtful in choosing the $\hat{f}_i$ when forming the average, $\hat{f}$. In particular we seek to minimize R. Lemma 2 tells how to do this. It says that if the entropies $\mathcal{E}_{\lambda_{\mathcal{L}_1}}(\hat{\mathcal{B}}_{\mathcal{L}_1})$, $\mathcal{E}_{\lambda_{\mathcal{L}_2}}(\hat{\mathcal{B}}_{\mathcal{L}_2})$ are close then the risks $\mathcal{R}\left(\hat{f}_1, f\right)$, $\mathcal{R}\left(\hat{f}_2, f\right)$ have a favorable, common upper bound. Here closeness will be measured in terms of the *entropy gap*, $\tau_{\mathcal{L}}$, of a given library $\mathcal{L}$, which is defined to be the difference between the entropy of the best (empirical) basis of that library and the entropy of the ideal basis:

$$\tau_{\mathcal{L}} = \left(\mathcal{E}_{\lambda_{\mathcal{L}}}(\mathcal{B}_{\mathcal{L}}) - \mathcal{E}_{\lambda_{\mathcal{L}}}(\hat{\mathcal{B}}_{\mathcal{L}})\right). \tag{14}$$

Recall the notation $\hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}}} = \hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}}}\left(y; \delta(\hat{\mathcal{B}}_{\mathcal{L}}, \lambda_{\mathcal{L}})\right)$.

Lemma 2. *Given two libraries $\mathcal{L}_i$, $i = 1, 2$, whose associated entropies satisfy:*

$$|\mathcal{E}_{\lambda_{\mathcal{L}_1}}(\hat{\mathcal{B}}_{\mathcal{L}_1}) - \mathcal{E}_{\lambda_{\mathcal{L}_2}}(\hat{\mathcal{B}}_{\mathcal{L}_2})| \leq \min_{k=1,2} \tau_{\mathcal{L}_k}, \tag{15}$$

then

$$\mathcal{R}\left(\hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}_i}}, f\right) = \mathbf{E}\left(||\hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}_i}} - f||^2\right) \leq A(\xi)\ \lambda_L \left(\sigma^2 + \min_{k=1,2} \mathcal{R}_{\mathcal{L}_k}(f)\right), \quad \textit{for } i = 1, 2. \tag{16}$$

Proof. Define the library $\bar{\mathcal{L}} = \{\hat{\mathcal{B}}_{\mathcal{L}_1}, \mathcal{B}_{\mathcal{L}_1}, \hat{\mathcal{B}}_{\mathcal{L}_2}, \mathcal{B}_{\mathcal{L}_2}\}$. We will apply Theorem 2 to $\bar{\mathcal{L}}$; to this end, we can take ξ in that theorem large enough in order to have $\lambda_{\bar{\mathcal{L}}} = \lambda_L$. Notice that by this choice of parameters the original entropies are unchanged, hence $\mathcal{E}_{\lambda_{\bar{\mathcal{L}}}}(\hat{\mathcal{B}}_{\mathcal{L}_i}) \leq \mathcal{E}_{\lambda_{\bar{\mathcal{L}}}}(\mathcal{B}_{\mathcal{L}_i})$ $i = 1, 2$. Without loss of generality assume $\mathcal{E}_{\lambda_L}(\hat{\mathcal{B}}_{\mathcal{L}_1}) \leq \mathcal{E}_{\lambda_L}(\hat{\mathcal{B}}_{\mathcal{L}_2})$. Then it follows from Theorem 2, applied to $\bar{\mathcal{L}}$, that in order to prove (16) we only need to prove

$$\mathbf{E}\left(||\hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}_2}} - f||^2\right) \leq A(\xi)\ \lambda_L \left(\sigma^2 + \mathcal{R}_{\mathcal{L}_1}(f)\right). \tag{17}$$

Construct now a new library $\mathcal{L}' = \{\mathcal{B}_{\mathcal{L}_1}, \hat{\mathcal{B}}_{\mathcal{L}_2}\}$; we will apply Theorem 2 to this library and will choose ξ in that theorem such that $\lambda_{\mathcal{L}'} = \lambda_L$. It follows from the hypothesis that $\hat{\mathcal{B}}_{\mathcal{L}'} = \hat{\mathcal{B}}_{\mathcal{L}_2}$. Equation (17) then follows from Theorem 2. □

Thus we should single out those libraries $\mathcal{L}_i$ which pairwise satisfy the hypothesis of Lemma 2. This gives a good value for R in equation (13) as evidenced by Proposition 1.

Let $|A|$ denote the cardinality of a set A.

Proposition 1. *Let $\mathcal{T} = \{\mathcal{L}_i | i = 1, \ldots, |\mathcal{T}|\}$, a collection of libraries, be given. If the associated entropies satisfy:*

$$|\mathcal{E}_{\lambda_{\mathcal{L}_i}}(\hat{\mathcal{B}}_{\mathcal{L}_i}) - \mathcal{E}_{\lambda_{\mathcal{L}_j}}(\hat{\mathcal{B}}_{\mathcal{L}_j})| \leq \min_{k=i,j} \tau_{\mathcal{L}_k} \quad i, j = 1, \ldots, \mathcal{T}, \tag{18}$$

then

$$\text{(19)}\quad \mathbf{E}\left(||(\frac{1}{|\mathcal{T}|}\sum_{i=1}^{|\mathcal{T}|}\hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}_i}}) - f||^2\right)$$

$$\leq A(\xi)\ \lambda_L\left(\sigma^2 + \min_{i=1,\dots,|\mathcal{T}|}\mathcal{R}_{\mathcal{L}_i}(f)\right) - \frac{1}{2\ |\mathcal{T}|^2}\left(\sum_{i=1}^{|\mathcal{T}|}\sum_{j=1}^{|\mathcal{T}|}\mathbf{E}\left(||\hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}_i}} - \hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}_j}}||^2\right)\right).$$

Proof. We will apply Lemma 1 to the following vectors

$$x_i = (\hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}_i}} - f)$$

and to the inner product $\langle g, h\rangle = \mathbf{E}\left(\sum_{i=1}^N g_i h_i\right)$ where g and h are random vectors i.e., functions defined in the probability space implicit in (1) and taking values in $\mathbf{R}^N$. Set R in Lemma 1 equal to

$$R^2 = A(\xi)\ \lambda_L\left(\sigma^2 + \min_{k=1,\dots,|\mathcal{T}|}\mathcal{R}_{\mathcal{L}_k}(f)\right).$$

The proof of (19) then follows from a direct application of Lemma 1 to the above setting and from inequality (20) below. We will prove,

$$\text{(20)}\quad ||x_i||^2 = \mathcal{R}\left(\hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}_i}}, f\right) \leq A(\xi)\ \lambda_L\left(\sigma^2 + \min_{k=1,\dots,|\mathcal{T}|}\mathcal{R}_{\mathcal{L}_k}(f)\right)\ \mathit{for}\ i = 1,\dots,|\mathcal{T}|.$$

We now prove (20) by induction on $p = |\mathcal{T}|$. The inductive hypothesis says that if (20) holds for any set $\mathcal{T}_{p-1} \subseteq \mathcal{T}$ with $|\mathcal{T}_{p-1}| = p - 1$ then it holds for $\mathcal{T}$ with $|\mathcal{T}| = p$. It follows from this inductive hypothesis, given the *pairwise* inequalities in (18), that in order to prove (20) we need to prove the statement only for $p = 2$. But this statement is simply Lemma 2. □

3.1. **Description of Shell Repeling Algorithm.** In this section we describe an algorithm which is based on Proposition 1. We work with a finite collection of many libraries $L = \{\mathcal{L}_i\}$. We assume that the best ortho-basis algorithm is applicable to each of the libraries $\mathcal{L}_i \in L$ and they satisfy the general properties listed at the beginning of Section 3. Let $\hat{\tau}_i = \hat{\tau}_{\mathcal{L}_i} = \hat{\tau}(y, \mathcal{L}_i, \sigma)$ be empirical estimates for $\tau_{\mathcal{L}_i} = (\mathcal{E}_{\lambda_L}(\mathcal{B}_{\mathcal{L}_i}) - \mathcal{E}_{\lambda_L}(\hat{\mathcal{B}}_{\mathcal{L}_i}))$. We will refer to these last two quantities as the estimated gaps and the true gaps respectively. The definition and computation of these quantities are deferred to Appendix A.

Algorithm: Without loss of generality we assume the best ortho bases $\hat{\mathcal{B}}_{\mathcal{L}}$ are indexed accordingly to increasing values of their entropies:

$$\text{(21)}\quad \mathcal{E}_{\lambda_L}(\hat{\mathcal{B}}_{\mathcal{L}_i}) \leq \mathcal{E}_{\lambda_L}(\hat{\mathcal{B}}_{\mathcal{L}_{i+1}})\ \mathit{for}\ i = 1,\dots,L.$$

From the collection $\{\hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}_i}} | \mathcal{L}_i \in L\}$ we will select a subset of these estimates which we will call the *repeling shell* and denote by $\mathcal{S}$. For simplicity, we will identify $\hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}}}$ with $\hat{\mathcal{B}}_{\mathcal{L}}$. The construction of $\mathcal{S}$ is done recursively as follows: set $\mathcal{S}^1 = \{\hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}_1}}\}$. Assume now that $\mathcal{S}^i$, $i \geq 1$ has been defined and set $p = |\mathcal{S}^i|$. Define

$$\text{(22)}\quad R_{\mathcal{S}^i} = \frac{1}{2\ p^2}\sum_{\hat{f}\in\mathcal{S}^i}\sum_{\hat{f}'\in\mathcal{S}^i}||\hat{f} - \hat{f}'||^2.$$

Inclussion Step: Use $\hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}_{i+1}}}$ to define

$$R_{i+1} = \frac{1}{(p+1)^2}\left(p^2 R_{\mathcal{S}^i} + \sum_{\hat{\mathcal{B}}_{\mathcal{L}} \in \mathcal{S}^i} ||\hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}_{i+1}}} - \hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}}}||^2\right). \tag{23}$$

Acceptance Step: then *if*

$$\mathcal{E}_{\lambda_L}(\hat{\mathcal{B}}_{\mathcal{L}_{i+1}}) - \mathcal{E}_{\lambda_L}(\hat{\mathcal{B}}_{\mathcal{L}_j}) \leq \min_{k=j,i+1} \hat{\tau}_k \; for\ all\ \hat{\mathcal{B}}_{\mathcal{L}_j} \in \mathcal{S}^i \tag{24}$$

and

$$R_{i+1} \geq R_{\mathcal{S}^i} \tag{25}$$

we let

$$\mathcal{S}^{i+1} = \mathcal{S}^i \cup \{\hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}_{i+1}}}\} \tag{26}$$

otherwise $\mathcal{S}^{i+1} = \mathcal{S}^i$ and we repeat above steps with $\hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}_{i+2}}}$.

The output of this algorithm is the following estimate

$$\hat{f}_{\mathcal{S}} = \frac{1}{|\mathcal{S}|}\sum_{i=1}^{|\mathcal{S}|} \hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}_i}} \tag{27}$$

The algorithm described above is clearly motivated by Proposition 1 which suggests to minimize the right hand side of (19). The algorithm does this by adding elements to the shell, hence making the first term of the right hand side of (19) smaller, if the "repeling" part becomes larger. We always include in the shell the estimate with smaller entropy this is to account for the cases in which one of the libraries is overwhelmingly better than the rest of the libraries to represent the underlying signal. An example is discussed in Section 4.

4. Numerical Examples

For our numerical examples we have used a collection of libraries, L, consisting of 41 different wavelet packet libraries $\mathcal{L}_i$ corresponding to given orthonormal wavelets including the Daubechies wavelets of orders 2–42 and 68 ([3] pg. 195), the Coiflets of order 6–30 ([3] pg. 198) and the symlets of order 8–30 ([3] pg. 261). If N in (1) is a power of 2, then each library has $M_L = N(\log_2(N)+1)$ distinct basis elements. We have taken $N = 2^{12}$. We define the Signal to Noise Ratio (SNR) of the data (1) to be

$$\mathrm{SNR}^2 = \frac{||f||_2^2}{\mathbf{E}\left(||\sigma\xi||_2^2\right)}.$$

In our numerical examples SNR = 6.0. For a given approximation, $\hat{f}$, of f we define the Root Mean Square Error (RMSE) to be

$$\frac{||f-\hat{f}||_2}{||f||_2}.$$

The numbers reported are obtained as averages over many random samples of noise. Similar numerical results are also obtained with different values of SNR and number of sample points.

Table 1 shows the best RMSE among the 41 best ortho basis estimates. This estimate, of course, is not available in practice and is of excellent quality. Table 1 also shows the average of the 41 RMSE. The RMSE of the 41 estimates varies considerably as can be seen from Table 3. Table 2 shows the RMSE for the shell repeling algorithm using the true gaps and estimated gaps.

Table 1. Best RMSE and average RMSE

Signal	Best RMSE	Average of RMSE
f^1	0.0325	0.0431
f^2	0.0386	0.0514
f^3	0.0148	0.0209
f^4	0.0404	0.0782
f^5	0.0452	0.0540
f^6	0.0294	0.0367

Table 2. Shell Repeling RMSE with True and Estimated Gaps

Signal	RMSE with True Gaps	RMSE with Estimated Gaps
f^1	0.0291	0.0293
f^2	0.0350	0.0350
f^3	0.0149	0.0144
f^4	0.0404	0.0530
f^5	0.0392	0.0388
f^6	0.0278	0.0286

The second column of the Table 3 shows the smallest value of entropy, which corresponds to the best empirical basis, among the 41 wavelets packets libraries. The third column shows the RMSE corresponding to that basis. The fourth column gives the value of entropy for the wavelet packet basis, that belongs to one of the 41 libraries, which gives the smallest RMSE value. This RMSE value is actually shown in column five. This table illustrates the shortcomings of choosing the estimate corresponding to the basis with smallest entropy among all the different libraries.

The numerical evidence reported indicates that the estimate of our algorithm, using either the true gaps or estimated gaps, is competitive with the best RMSE estimate. The case of f^4 deserves special discussion. This "block function" is best reconstructed with the Haar wavelet packet library (included among our 41 libraries). In this case, the shell $\mathcal{S}$, with the true gaps, actually consists of only one element, which is the estimate corresponding the the Haar library. The reason being that the true gap for that library turns out to be very small relative to the remaining set of libraries. On the other hand, our estimates $\hat{\tau}_{\mathcal{L}}$ tend to overestimate the true gaps, this is the reason for the poorer performance of our algorithm in this case.

Table 3. Empirical Entropies and RMSE for all signals.

	Best Empirical Basis		Best RMSE Basis	
Signals	Entropy (smallest)	RMSE	Entropy	RMSE (smallest)
f^1	0.0278	0.0997	0.0293	0.0325
f^2	0.0278	0.0868	0.0285	0.0386
f^3	0.0274	0.0328	0.0279	0.0148
f^4	0.0278	0.0406	0.0278	0.0404
f^5	0.0283	0.0867	0.0288	0.0452
f^6	0.0278	0.0441	0.0280	0.0294

5. Discussion and Extensions

Our paper exploits some aspects of Donoho and Johnstone's framework, namely the convexity of the L^2 norm and the oracle inequalities. We present a framework that justifies the use of averages when many libraries are present and gives indication under what conditions these averages give better estimates. The observation that bases with entropy sufficiently close can not be distinguished from the point of view of oracle inequalities lead us to propose the algorithms in Section 3.1. An alternative to this approach, which is presently being investigated by the authors, is to construct the shell $\mathcal{S}$ by means of an optimization problem. This approach allows, in particular, for more than one basis from a given library to appear in the final average estimate. This type of approach requires a new type of oracle inequality where the risk functional on the right hand side of equation (12) is replaced for the ideal risk of an average obtained through the help of an oracle.

Appendix A. Entropy Gaps for Libraries of Wavelet Packets

We remark that in order to compute the quantity $\mathcal{E}_{\mathcal{L}}(\mathcal{B}_{\mathcal{L}})$ we need the values of the *noisy* coefficients $\theta_i(y, \mathcal{B}_{\mathcal{L}})$. That is, knowledge of the underlying function is needed only to compute $\mathcal{B}_{\mathcal{L}}$. We have found, from extensive numerical experimentations, that the following estimates are useful estimates for the true gaps. As before, let L be the collection of wavelet packets under consideration. A simple computation shows that the ideal basis satisfies

$$\mathcal{B}_{\mathcal{L}} = \arg\min_{\mathcal{B}\in\mathcal{L}} \sum_{i=1}^{N} \min\left(\theta_i^2(f, \mathcal{B}), \sigma^2\right). \tag{28}$$

Moreover

$$\mathcal{E}_{\lambda_{\mathcal{L}}}(\mathcal{B}_{\mathcal{L}}) = \sum_{i=1}^{N} \min\left(\theta_i^2(y, \mathcal{B}_{\mathcal{L}}), \sigma^2\, \lambda_{\mathcal{L}}\right). \tag{29}$$

The idea is to replace f above by its estimates and then take averages over the corresponding entropies. Precisely, define

$$\mathcal{B}_{\mathcal{L}}^{'} = \mathcal{B}_{\mathcal{L}}(\hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}'}}) = \arg\min_{\mathcal{B}\in\mathcal{L}} \sum_{i=1}^{N} \min\left(\theta_i^2(\hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}'}}, \mathcal{B}), \sigma^2\right) \tag{30}$$

and the corresponding entropy

$$\mathcal{E}_{\lambda_{\mathcal{L}}}(\mathcal{B}^{'}_{\mathcal{L}}) = \sum_{i=1}^{N} \min\left(\theta_i^2(y, \mathcal{B}^{'}_{\mathcal{L}}), \sigma^2\, \lambda_{\mathcal{L}}\right). \tag{31}$$

Finally

$$\hat{\mathcal{E}}_{\lambda_{\mathcal{L}}} = \frac{1}{|L|} \sum_{\mathcal{L}' \in L} \mathcal{E}_{\lambda_{\mathcal{L}}}(\hat{f}_{\hat{\mathcal{B}}_{\mathcal{L}'}}) \tag{32}$$

and

$$\hat{\tau}_i = \hat{\tau}_{\mathcal{L}_i} = \hat{\tau}(y, \mathcal{L}_i, \sigma) = \hat{\mathcal{E}}_{\lambda_{\mathcal{L}}} - \mathcal{E}_{\lambda_{\mathcal{L}}}(\hat{\mathcal{B}}_{\mathcal{L}}) \tag{33}$$

Because $\mathcal{B}^{'}_{\mathcal{L}} \in \mathcal{L}$ we have $\hat{\tau}_i \geq 0$.

A.1. **Signals Used In Numerical Experiments.** Formulas for two of the six signals are given below. Four of the six functions are borrowed from [6]: f^2 is Doppler, f^4 is Blocks, f^5 is Bumps and f^6 is Heavisine. In practice the data in (1) represents N equally spaced samples of a given function on the closed interval from 0 to 1. For convenience each set of sample points is normalized so as to have l^2 norm equal to one.

f^1 is the function,

$$f(t) = t^2(1-t)^2\cos(200t^2),$$

and f^3 is the function

$$f(t) = t(t - \frac{1}{2})^2(t-1)^3.$$

References

[1] R.R. Coifman, Y. Meyer, Remarques sur l'analyse de Fourier à fenêtre. *C. R. Acad. Sci. Paris, Série I* **312** (1991) 259-261.

[2] R.R. Coifman, M.V. Wickerhauser, Entropy-based algorithms for best basis selection. *IEEE transactions on Information Theory* **38** (1992) 712-718.

[3] I. Daubechies, *Ten Lectures on Wavelets* (SIAM Press, Philedelphia, 1992).

[4] D. L. Donoho, CART and Best-OrthoBasis: a connection *Preprint.*

[5] D. L. Donoho, I.M. Johnstone, Ideal denoising in an orthonormal basis chosen from a library of bases. *Comptes Rendus Aca. Sci. Paris A* **319** (1994) 1327-1322.

[6] D. L. Donoho, I.M. Johnstone, Ideal spatial adaptation by wavelet shrinkage. *Biometrica* **81** (1994) 425-455.

[7] D. L. Donoho, I.M. Johnstone, Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Stat. Assn.* **90** (1995) 1200-1224.

[8] V. Wickerhauser, *Adapted Wavalet Analysis from Theory to Software* (A.K. Peters, Wellesley, 1994).

Department of Mathematics, Physics and Computer Science, Ryerson Polytechnic University, Toronto, Ontario M5B 2K3, Canada.

E-mail address: ferrando@acs.ryerson.ca

E-mail address: lkolasa@acs.ryerson.ca