

Nonlinear Wavelet Image Processing: Variational Problems, Compression, and Noise Removal through Wavelet Shrinkage*

ANTONIN CHAMBOLLE¹, RONALD A. DEVORE², NAM-YONG LEE³, AND BRADLEY J. LUCIER⁴

Abstract

This paper examines the relationship between wavelet-based image processing algorithms and variational problems. Algorithms are derived as exact or approximate minimizers of variational problems; in particular, we show that wavelet shrinkage can be considered the exact minimizer of the following problem: given an image F defined on a square I , minimize over all g in the Besov space $B_1^1(L_1(I))$ the functional $\|F - g\|_{L_2(I)}^2 + \lambda \|g\|_{B_1^1(L_1(I))}$. We use the theory of nonlinear wavelet image compression in $L_2(I)$ to derive accurate error bounds for noise removal through wavelet shrinkage applied to images corrupted with i.i.d., mean zero, Gaussian noise. A new signal-to-noise ratio, which we claim more accurately reflects the visual perception of noise in images, arises in this derivation. We present extensive computations that support the hypothesis that near-optimal shrinkage parameters can be derived if one knows (or can estimate) only two parameters about an image F : the largest α for which $F \in B_q^\alpha(L_q(I))$, $1/q = \alpha/2 + 1/2$, and the norm $\|F\|_{B_q^\alpha(L_q(I))}$. Both theoretical and experimental results indicate that our choice of shrinkage parameters yields uniformly better results than Donoho and Johnstone's VisuShrink procedure; an example suggests, however, that Donoho and Johnstone's SureShrink method, which uses a different shrinkage parameter for each dyadic level, achieves lower error than our procedure.

1. Introduction

This paper has several objectives. The first is to describe several families of variational problems that can be solved quickly using wavelets. These variational problems take the form: given a positive parameter λ and an image, a signal, or noisy data $f(x)$ defined for x in some finite domain I , find a function \tilde{f} that minimizes over all possible functions g the functional

$$(1) \quad \|f - g\|_{L_2(I)}^2 + \lambda \|g\|_Y,$$

where

$$\|f - g\|_{L_2(I)} := \left(\int_I |f(x) - g(x)|^2 dx \right)^{1/2}$$

is the root-mean-square *error* (or more generally, difference) between f and g , and $\|g\|_Y$ is the norm of the approximation g in a *smoothness space* Y . The original image f could be noisy, or it could simply be “messy” (a medical image, for example), while \tilde{f} would be a denoised, segmented, or compressed version of f . The amount of noise removal, compression, or segmentation is determined by the parameter λ ; if λ is large, then necessarily $\|g\|_Y$ must be smaller at the minimum, i.e., g must be smoother, while when λ is small, g can be rough, with $\|g\|_Y$ large, and one achieves a small error at the minimum.

These types of variational problems have become fairly common in image processing and statistics; see, e.g., [32]. For example, Rudin-Osher-Fatemi [33] set Y to the space of functions of bounded variation $BV(I)$ for images (see also [1]), and non-parametric estimation sets Y to be the Sobolev space $W^m(L_2(I))$ of functions all of whose m th derivatives are square-integrable; see the monograph by Wahba [34]. In fact, Y can be very general; one could, for example, let Y contain all piecewise constant functions, with $\|g\|_Y$ equal to the number of different pieces or segments of g ; this would result in segmentation of the original image f . Indeed, Morel and Solimini [31] argue that almost any reasonable segmentation algorithm can be posed in this form. Techniques like this are also known as Tikhonov regularization; see [2]. In [12] we considered (1) in the context of interpolation of function spaces; in that theory, the infimum of (1) over all g is $K(f, \lambda, L_2(I), Y)$, the K -functional of f between $L_2(I)$ and Y .

A fast way of solving (1) is required for practical algorithms. In [12], we noted that the norms of g in many function spaces Y can be expressed in terms of the wavelet coefficients of g . In other words, if we choose an (orthogonal or biorthogonal) wavelet basis for $L_2(I)$, and we expand g in terms of its wavelet coefficients, then the norm $\|g\|_Y$ is equivalent to a norm of the wavelet coefficients of g ; see, for example, [30], [14], or [23].

In [12] we proposed that by choosing $\|g\|_Y$ to be one of these norms and by calculating approximate minimizers rather than exact minimizers, one can find efficient computational algorithms in terms of the wavelet coefficients of the data f . In particular, we showed how choosing $Y = W^m(L_2(I))$ and approximately minimizing (1) leads to wavelet algorithms that are analogous to well-known linear algorithms for compression and noise removal. In particular, we find that the wavelet coefficients of \tilde{f} are simply all wavelet coefficients of f with frequency below a fixed value, determined by λ .

Additionally, we proposed choosing Y from the family

* A shorter version of this paper appeared in the *IEEE Transactions on Image Processing*, v. 7, 1998, pp. 319–335.

¹CEREMADE (CNRS URA 749), Université de Paris–Dauphine, 75775 Paris CEDEX 16, France, Antonin.Chambolle@ceremade.dauphine.fr. Supported by the CNRS.

²Department of Mathematics, University of South Carolina, Columbia, SC 29208, devore@math.sc.edu. Supported in part by the Office of Naval Research, Contract N00014-91-J-1076.

³Department of Mathematics, Purdue University, West Lafayette, IN 47907-1395, nylee@math.purdue.edu. Supported in part by the Purdue Research Foundation.

⁴Department of Mathematics, Purdue University, West Lafayette, IN 47907-1395, lucier@math.purdue.edu. Supported in part by the Office of Naval Research, Contract N00014-91-J-1152. Part of this work was done while the author was a visiting scholar at CEREMADE, Université de Paris–Dauphine, Paris, France.

of Besov spaces $B_q^\alpha(L_p(I))$, with $\alpha > 0$ and $q = p$ satisfying

$$(2) \quad 1/q = \alpha/2 + 1/2.$$

These spaces, which contain, roughly speaking, functions with α derivatives in $L_q(I)$, arise naturally in two contexts. First, in image compression using wavelets, if f can be approximated to $O(N^{-\alpha/2})$ in $L_2(I)$ by wavelet sums with N nonzero terms, then f is necessarily in $B_q^\alpha(L_q(I))$. (Because of certain technicalities, this statement only approximates the truth; see [6] for precise statements and definitions.) Conversely, if f is in $B_q^\alpha(L_q(I))$, then scalar quantization of the wavelet coefficients with scale-dependent quantization levels yields compression algorithms with convergence rates of $O(N^{-\alpha/2})$ in $L_2(I)$. (There is a complete theory for compression in $L_p(I)$ for $0 < p < \infty$ [8] and $p = \infty$ [7] and [13].) We emphasize that this is an *equivalence*—you achieve a given rate of approximation with wavelet image compression *if and only if* f is in the corresponding Besov smoothness class. Second, one can ask which Besov spaces $B_q^\alpha(L_q(I))$ of *minimal smoothness* are embedded in $L_2(I)$. One wishes to use function spaces of minimal smoothness to allow as many sample functions g as possible. A variant of the Sobolev embedding theorem implies that Besov spaces of minimal smoothness necessarily satisfy (2). We note that restricting attention to Besov spaces with $p = q$ does not allow us to consider other important spaces, such as $BV(I)$ as proposed by Rudin-Osher-Fatemi [33], or the space $B_2^1(L_1(I))$; both these spaces, while slightly larger than $B_1^1(L_1(I))$ (which does satisfy (2)), are also contained in $L_2(I)$.

When $Y = B_q^\alpha(L_q(I))$, approximate minimizers \tilde{f} of (1) have wavelet expansions containing the wavelet coefficients of f larger than a threshold determined by λ . These nonlinear algorithms are related to threshold coding or certain types of progressive transmission in image compression. In [12], we provided simple analyses of the performance of these algorithms. In all cases, the nonlinear algorithms are preferable to the linear for two reasons. First, the nonlinear algorithm achieves a given performance level for more images than the linear algorithm does. Second, for a fixed image, it is possible (even likely) that the nonlinear algorithm will achieve a higher level of performance than the linear algorithm; the converse never occurs in compression.

These nonlinear algorithms are related to the large body of work by Donoho and Johnstone on what they call *wavelet shrinkage*. Indeed, whereas the nonlinear algorithms derived in [12] threshold the wavelet coefficients of f to find the coefficients of \tilde{f} , wavelet shrinkage takes the coefficients of absolute value larger than the threshold and shrinks them by the threshold value towards zero. We show in Section 3 that wavelet shrinkage is the exact minimizer of (1) when $Y = B_1^1(L_1(I))$ and $\|g\|_{B_1^1(L_1(I))}$ is given by a (wavelet-dependent) norm equivalent to the usual $\|g\|_{B_1^1(L_1(I))}$ norm. (Examining the case $Y = B_1^1(L_1(I))$ in (2) was actually motivated by the practical success of set-

ting $Y = BV(I)$ in [33].) In a series of papers (see, e.g., [19] and [21]), Donoho and Johnstone show that wavelet shrinkage leads to near-optimal noise removal properties when the images f are modeled stochastically as members of several Besov spaces $B_q^\alpha(L_p(I))$.

The second goal of this paper is to prove claims presented in [12] about the rate of Gaussian noise removal from images. We continue the program of [6] and [12] in advocating a deterministic smoothness model for images (see also [29]). This model has proved highly successful in image compression [6], and we show here that it also leads to good results in analyzing noise removal. This model assigns two numbers, a smoothness class specified by α , and a norm $\|f\|_{B_q^\alpha(L_q(I))}$, $1/q = \alpha/2 + 1/2$, to each image, and uses these numbers to choose the parameter λ . We show that knowing (or estimating) α leads to a better estimate of the smoothing parameter, and knowing $\|f\|_{B_q^\alpha(L_q(I))}$ allows one to make an even finer estimate of the optimal λ . (Note: After this paper was submitted, we discovered that Donoho and Johnstone [20] have previously calculated the same first-order dependence of the error on α ; in some ways, our arguments parallel theirs.)

Our results have several properties that make them useful for practical image processing. We consider images in Besov spaces of minimal smoothness $B_q^\alpha(L_q(I))$ with $1/q = \alpha/2 + 1/2$. If images are assumed to have less smoothness, then weaker (or no) results hold. We explicitly consider models of data and observations in which measurements are not point values, but measurements of integrals of the image with local point spread functions. (One can think of this as point evaluation of a convolution of the image with a smoothing kernel.) The measurement functionals can be averages of an intensity field F over pixels in the image, a model that closely fits the physics of CCD cameras. Not using point values is a mathematical necessity, because point values are not defined for the spaces of minimal smoothness we consider, and are not well defined for images. (One cannot define the point value of an image at an internal edge separating regions of light and dark, for example.) Restricting the data acquisition model to point evaluation implies that images are continuous (and that $1/q \leq \alpha/2$), which may be natural in some contexts, but not for image processing, in which intensity fields are more naturally modeled as discontinuous functions.

The final goal of this paper is to provide rather sharp estimates of the best wavelet shrinkage parameter in removing Gaussian noise from images. We show through rather extensive computational examples that our analysis often leads to a λ that is within 10% of the optimal λ , and which is generally 1/2 to 1/4 the shrinkage parameter suggested by Donoho and Johnstone in their VisuShrink method. In other words, the VisuShrink parameter leads to oversmoothing of the noisy image, resulting in the unnecessary loss of image details.

The rest of the paper is organized as follows. In Section 2, we review some properties of wavelets and smooth-

ness spaces that we need in the following sections. In Section 3, we recall briefly from [12] how our abstract framework leads quite naturally to common algorithms in image processing, and we expand further on this method by solving several more variational problems of interest. It is here that we show that wavelet shrinkage is equivalent to solving (1) with $Y = B_1^1(L_1(I))$. In Section 4, we show how to compute accurate wavelet-based image representations given pixel measurements of very general form. Donoho discusses another approach to this problem in [16]. In Section 5, we review the simplified theory of wavelet compression in $L_2(I)$ presented in [12]; we use this theory in Section 8 on noise removal. In Section 6, we show that quantization strategies for optimal $L_2(I)$ compression lead to optimal image compression when the error is measured in $B_r^\beta(L_r(I))$, $1/r = \beta/2 + 1/2$, for $0 < \beta < \alpha$, when the image is in $B_q^\alpha(L_q(I))$, $1/q = \alpha/2 + 1/2$. In Section 7, we argue that the perception by the Human Visual System (HVS) of error induced by various image processing tasks is strongly influenced not only by the difference between two images, but also by changes in local smoothness. In Section 8, we formulate our noise-removal model and prove the main result of the paper on the rate of noise removal by wavelet shrinkage. It is here that we introduce a new signal-to-noise ratio that is useful in estimating the size of the shrinkage parameter. In Section 9, we discuss why the preceding results apply not only to computations with orthogonal wavelets, but also to computations with biorthogonal wavelets. Donoho presents a similar analysis in [17]. In Section 10, we present the results of rather extensive computational tests of wavelet shrinkage applied with the shrinkage parameter suggested in Section 8. Although our results are uniformly better than the VisuShrink procedure of Donoho and Johnstone [19], it appears from a single example that their later SureShrink procedure [18], which uses different shrinkage parameters for different dyadic levels, may give better results than our method in some cases. Finally, in the Appendices, we provide proofs for several statements in Sections 3 and 4.

Authors' notes. We take the opportunity here to remark on issues arising from several previous papers.

Extra assumptions are needed in [12] to prove the stated rate of convergence of the noise removal algorithm with oracle; one can assume, for example, that the image intensity is bounded, which is quite natural for problems in image processing.

In [10], we presented an interpretation of certain bi-orthogonal wavelets as derived in [3] and [25] that owed much to the connection between these wavelets and function reconstruction from cell averages as used in computational fluid dynamics. This connection was first recognized and developed by Ami Harten (see, e.g., [24]), whose work provided the inspiration for the approach taken in [10]; we regret that Harten's work was not properly recognized in [10].

2. Wavelets and Smoothness Spaces

In this paper, images (light intensity fields) are functions f defined on the square $I := [0, 1]^2$, and we consider variational problems of the form

$$\min_g \{ \|f - g\|_{L_2(I)}^2 + \lambda \|g\|_Y^s \}$$

where Y is a space of test functions (generally embedded in $L_2(I)$), λ is a positive parameter, and s is an exponent that is chosen to make the computations (and analysis) easier.

In [12] we suggested using spaces Y for which the norm of g in Y is equivalent to a sequence norm of the wavelet coefficients of g . Let us first consider (real) orthogonal wavelets on I as described by Cohen, Daubechies, and Vial [4]. One begins with a one-dimensional orthogonal wavelet ψ , such that if we set $\psi_{j,k}(x) := 2^{k/2}\psi(2^kx - j)$ to be the scaled (by $2^{k/2}$) translated (by $j/2^k$) dilates (by 2^k) of the original ψ , then $\{\psi_{j,k}\}_{j,k \in \mathbb{Z}}$ forms an orthonormal basis for $L_2(\mathbb{R})$, that is, for the coefficients $c_{j,k} := \int_{\mathbb{R}} f(x)\psi_{j,k}(x) dx$

$$f = \sum_{j,k \in \mathbb{Z}} c_{j,k} \psi_{j,k} \quad \text{and} \quad \|f\|_{L_2(\mathbb{R})}^2 = \sum_{j,k \in \mathbb{Z}} c_{j,k}^2.$$

Associated with ψ is a scaling function ϕ , from which one generates the functions $\phi_{j,k}(x) := 2^{k/2}\phi(2^kx - j)$. The set $\{\phi_{j,k}\}_{j \in \mathbb{Z}}$ is orthonormal for fixed k . For example, the Haar wavelets have $\phi = \chi_{[0,1]}$, the characteristic function of the interval $[0, 1]$, and $\psi = \chi_{[0,1/2]} - \chi_{[1/2,1]}$. We can easily construct two-dimensional wavelets from the one-dimensional ψ and ϕ by setting for $x := (x_1, x_2) \in \mathbb{R}^2$

$$\begin{aligned} \psi^{(1)}(x_1, x_2) &:= \psi(x_1)\phi(x_2), \quad \psi^{(2)}(x_1, x_2) := \phi(x_1)\psi(x_2), \\ &\text{and } \psi^{(3)}(x_1, x_2) := \psi(x_1)\psi(x_2). \end{aligned}$$

If we let $\Psi := \{\psi^{(1)}, \psi^{(2)}, \psi^{(3)}\}$, then the set of functions $\{\psi_{j,k}(x) := 2^{k/2}\psi(2^kx - j)\}_{\psi \in \Psi, k \in \mathbb{Z}, j \in \mathbb{Z}^2}$ forms an orthonormal basis for $L_2(\mathbb{R}^2)$, i.e., for every $f \in L_2(\mathbb{R}^2)$ there are coefficients $c_{j,k,\psi} := \int_{\mathbb{R}^2} f(x)\psi_{j,k}(x) dx$ such that

$$f = \sum_{\substack{j \in \mathbb{Z}^2, k \in \mathbb{Z}, \\ \psi \in \Psi}} c_{j,k,\psi} \psi_{j,k} \quad \text{and} \quad \|f\|_{L_2(\mathbb{R}^2)}^2 = \sum_{\substack{j \in \mathbb{Z}^2, k \in \mathbb{Z}, \\ \psi \in \Psi}} c_{j,k,\psi}^2.$$

Instead of considering the sum over all dyadic levels k , one can sum over $k \geq K$ for a fixed K ; in this case, we have

$$\begin{aligned} f &= \sum_{\substack{j \in \mathbb{Z}^2, k \geq K, \\ \psi \in \Psi}} c_{j,k,\psi} \psi_{j,k} + \sum_{j \in \mathbb{Z}^2} d_{j,K} \phi_{j,K} \quad \text{and} \\ \|f\|_{L_2(\mathbb{R}^2)}^2 &= \sum_{\substack{j \in \mathbb{Z}^2, k \geq K, \\ \psi \in \Psi}} c_{j,k,\psi}^2 + \sum_{j \in \mathbb{Z}^2} d_{j,K}^2, \end{aligned}$$

where $d_{j,K} = \int_{\mathbb{R}^2} f(x)\phi_{j,K}(x) dx$.

When one is concerned with a finite domain, e.g., the square I , then two changes must be made to this basis for all of $L_2(\mathbb{R}^2)$ to obtain an orthonormal basis for $L_2(I)$. First, one does not consider all scales $k \in \mathbb{Z}$, but only non-negative scales $k \geq 0$, and not all shifts $j \in \mathbb{Z}^2$, but only

those shifts for which $\psi_{j,k}$ intersects I nontrivially. Second, one must adapt the wavelets that overlap the boundary of I in order to preserve orthogonality on the domain. (Specifically, the modified $\psi_{j,k}$ for $k = 0$ look more like the functions $\phi_{j,k}$, $k = 0$.) There are several ways to do this; the paper [4] gives perhaps the best way and some historical comparisons. To ignore all further complications of this sort, we shall not precisely specify the domains of the indices of the sums and write for $f \in L_2(I)$

$$(3) \quad f = \sum_{j,k,\psi} c_{j,k,\psi} \psi_{j,k} \quad \text{and} \quad \|f\|_{L_2(I)}^2 = \sum_{j,k,\psi} c_{j,k,\psi}^2.$$

Not only can one determine whether f is in $L_2(I)$ by examining the coefficients $\{c_{j,k,\psi}\}$, but one can also determine whether f is in many different function spaces Y . We shall consider the family of Besov spaces $B_q^\alpha(L_p(I))$, $0 < \alpha < \infty$, $0 < p \leq \infty$, and $0 < q \leq \infty$. These spaces have, roughly speaking, α “derivatives” in $L_p(I)$; the third parameter q allows one to make finer distinctions in smoothness. Various settings of the parameters yield more familiar spaces. For example, when $p = q = 2$, then $B_2^\alpha(L_2(I))$ is the Sobolev space $W^\alpha(L_2(I))$, and when $\alpha < 1$, $1 \leq p \leq \infty$, and $q = \infty$, $B_\infty^\alpha(L_p(I))$ is the Lipschitz space $\text{Lip}(\alpha, L_p(I))$.

When $p < 1$ or $q < 1$, then these spaces are not complete normed linear spaces, or Banach spaces, but rather complete quasi-normed linear spaces; that is, the triangle inequality may not hold, but for each space $B_q^\alpha(L_p(I))$ there exists a constant C such that for all f and g in $B_q^\alpha(L_p(I))$,

$$\|f + g\|_{B_q^\alpha(L_p(I))} \leq C(\|f\|_{B_q^\alpha(L_p(I))} + \|g\|_{B_q^\alpha(L_p(I))}).$$

With a certain abuse of terminology, we shall continue to call these quasi-norms norms.

The Besov space norm can be defined intrinsically in terms of moduli of smoothness. We give the definition here for the interested reader.

For any $h \in \mathbb{R}^2$, we define $\Delta_h^0 f(x) := f(x)$ and

$$\Delta_h^{k+1} f(x) := \Delta_h^k f(x+h) - \Delta_h^k f(x), \quad k = 0, 1, \dots$$

For $r > 0$, $\Delta_h^r f(x)$ is defined for $x \in I_{rh} := \{x \in I \mid x + rh \in I\}$. The $L_p(I)$ -modulus of smoothness, $0 < p \leq \infty$, is defined as

$$\omega_r(f, t)_p := \sup_{|h| \leq t} \left(\int_{I_{rh}} |\Delta_h^r f(x)|^p dx \right)^{1/p},$$

with the usual change to an essential supremum when $p = \infty$. Given $\alpha > 0$, $0 < p \leq \infty$ and $0 < q \leq \infty$, choose $r \in \mathbb{Z}$ with $r > \alpha \geq r - 1$. Then the the Besov space seminorm is defined as

$$|f|_{B_q^\alpha(L_p(I))} := \left(\int_0^\infty [t^{-\alpha} \omega_r(f, t)_p]^q \frac{dt}{t} \right)^{1/q},$$

again with a supremum when $q = \infty$. The Besov space norm is

$$\|f\|_{B_q^\alpha(L_p(I))} = |f|_{B_q^\alpha(L_p(I))} + \|f\|_{L_p(I)}.$$

The application of Besov spaces to image compression with wavelets can be found in [6]; we need here only the following facts. Assume that α and p satisfy $1/p < \alpha/2 + 1$, so that $B_q^\alpha(L_p(I))$ is embedded in $L_1(I)$. If there exists an integer $r > \alpha$ such that for all $\psi \in \Psi$ and all pairs of nonnegative integers $s = (s_1, s_2)$ with $|s| = s_1 + s_2$ and $x^s = x_1^{s_1} x_2^{s_2}$,

$$\int_I x^s \psi(x) dx = 0 \quad \text{for } |s| < r$$

and $\psi \in B_q^\beta(L_p(I))$ for some $\beta > \alpha$ (the set of β and α for which this is true depends on Ψ), then the norm $\|f\|_{B_q^\alpha(L_p(I))}$ is equivalent to a norm of the sequence of coefficients $\{c_{j,k,\psi}\}$,

$$(4) \quad \|f\|_{B_q^\alpha(L_p(I))} \asymp \left(\sum_k \left(\sum_{j,\psi} 2^{\alpha k p} 2^{k(p-2)} |c_{j,k,\psi}|^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}}.$$

When $p = q$, there is an obvious simplification

$$(5) \quad \|f\|_{B_p^\alpha(L_p(I))} \asymp \left(\sum_k \sum_{j,\psi} 2^{\alpha k p} 2^{k(p-2)} |c_{j,k,\psi}|^p \right)^{1/p}.$$

In this paper $A(f) \asymp B(f)$ means that there exist positive constants C_1 and C_2 such that for all f , $C_1 A(f) \leq B(f) \leq C_2 A(f)$. The constants C_1 and C_2 depend on the parameters α , p , and q , and on the wavelet basis $\{\psi_{j,k}\}$; the expression on the right of (5) is wavelet dependent.

We always use the equivalent sequence norm (4) in our calculations with $\|f\|_{B_q^\alpha(L_p(I))}$.

In the variational problem (1) the difference between f and g is always measured in $L_2(I)$. Thus, there are two scales of Besov spaces of importance. The first is $B_2^\alpha(L_2(I)) = W^\alpha(L_2(I))$, which measures smoothness of order α in $L_2(I)$, for which

$$\|f\|_{B_2^\alpha(L_2(I))} \asymp \left(\sum_k \sum_{j,\psi} 2^{2\alpha k} |c_{j,k,\psi}|^2 \right)^{1/2}.$$

The second is the scale of spaces $B_q^\alpha(L_q(I))$ with $1/q = \alpha/2 + 1/2$; these are the spaces of the form $B_p^\alpha(L_q(I))$ with $p = q$ of minimal smoothness to be embedded in $L_2(I)$, for which

$$\|f\|_{B_q^\alpha(L_q(I))} \asymp \left(\sum_k \sum_{j,\psi} |c_{j,k,\psi}|^q \right)^{1/q}.$$

Another important fact that arises immediately from (5) is that $B_p^\alpha(L_p(I))$ is embedded in $L_2(I)$ if $\alpha p + p - 2 \geq 0$, or $1/p \leq \alpha/2 + 1/2$; see, e.g., [30].

We need a bound on the smoothness in the Sobolev space $B_2^{\alpha/(\alpha+1)}(L_2(I)) = W^{\alpha/(\alpha+1)}(L_2(I))$ of bounded functions in $B_q^\alpha(L_q(I))$, $1/q = \alpha/2 + 1/2$. Our argument is typical of those used in the theory of interpolation of function spaces. For any bounded F (not a practical restriction for images), we have

$$\begin{aligned} |c_{j,k,\psi}| &= |\langle F, \psi_{j,k} \rangle| \leq \|F\|_{L_\infty(I)} \|\psi_{j,k}\|_{L_1(I)} \\ &\leq C 2^{-k} \|F\|_{L_\infty(I)}. \end{aligned}$$

It follows that F is in $B_2^{\alpha/(\alpha+1)}(L_2(I))$, since

$$\begin{aligned} \|F\|_{B_2^{\alpha/(\alpha+1)}(L_2(I))}^2 &= \sum_{k,j,\psi} 2^{2k\alpha/(\alpha+1)} |c_{j,k,\psi}|^2 \\ &= \sum_{k,j,\psi} 2^{2k\alpha/(\alpha+1)} |c_{j,k,\psi}|^{2-q} |c_{j,k,\psi}|^q \\ &\leq C \|F\|_{L_\infty(I)}^{2-q} \sum_{k,j,\psi} 2^{2k\alpha/(\alpha+1)} 2^{-(2-q)k} |c_{j,k,\psi}|^q \\ &= C \|F\|_{L_\infty(I)}^{2-q} \sum_{k,j,\psi} |c_{j,k,\psi}|^q, \end{aligned}$$

since $2 - q = 2\alpha/(\alpha + 1)$. Thus,

$$(6) \quad \|F\|_{B_2^{\alpha/(\alpha+1)}(L_2(I))} \leq C \|F\|_{L_\infty(I)}^{\alpha/(\alpha+1)} \|F\|_{B_q^\alpha(L_q(I))}^{1/(\alpha+1)}.$$

3. Solving Variational Problems with Wavelets

Following the suggestion in [12], we now consider the problem: Find \tilde{f} that minimizes over all g the functional

$$(7) \quad \|f - g\|_{L_2(I)}^2 + \lambda \|g\|_{B_p^\alpha(L_p(I))}^p.$$

Using (3) and (5), we can expand f and g in their wavelet expansions

$$f = \sum_{j,k,\psi} c_{j,k,\psi} \psi_{j,k} \quad \text{and} \quad g = \sum_{j,k,\psi} d_{j,k,\psi} \psi_{j,k},$$

and solve instead the equivalent problem of finding the minimizer of the functional

$$(8) \quad \sum_{j,k,\psi} |c_{j,k,\psi} - d_{j,k,\psi}|^2 + \lambda \sum_{j,k,\psi} 2^{\alpha k p} 2^{k(p-2)} |d_{j,k,\psi}|^p.$$

One notes immediately that the infinite-dimensional non-linear problem (7) completely decouples in the wavelet representation to the separable problem (8). That is, one minimizes (8) by minimizing separately over $d_{j,k,\psi}$

$$|c_{j,k,\psi} - d_{j,k,\psi}|^2 + \lambda 2^{\alpha k p} 2^{k(p-2)} |d_{j,k,\psi}|^p$$

for each j, k , and ψ .

While (8) can be minimized exactly in several interesting cases, an approximate minimizer can always be found. The problem reduces to finding the minimizer s , given t , of

$$(9) \quad E(s) := |s - t|^2 + \mu |s|^p,$$

where $t = c_{j,k,\psi}$ and $\mu = \lambda 2^{\alpha k p} 2^{k(p-2)}$. First, note that if s is not between 0 and t , then we can reduce $E(s)$ by changing s to be the closer of 0 and t ; thus, we can assume without loss of generality that s is between 0 and t .

Next, we remark that if $|s| \leq |t|/2$, then $E(s)$ is no less than $|t|^2/4$, and if $|s| \geq |t|/2$, then $E(s)$ is no less than $\mu |t|^p / 2^p$. Thus, if we set

$$s = \begin{cases} 0, & |t|^2 \leq \mu |t|^p, \\ t, & |t|^2 \geq \mu |t|^p, \end{cases}$$

we have $E(s) = \min(|t|^2, \mu |t|^p)$, which is within a factor of $\max(4, 2^p)$ of the minimum of (9)

Using this formula for s , we can construct an approximate minimizer $\tilde{f} = \sum_{j,k,\psi} \tilde{c}_{j,k,\psi} \psi_{j,k}$ to (7). In the special case $p = 2$, this reduces to setting $s = t$ when $\mu \leq 1$, i.e.,

$$(10) \quad \lambda 2^{2\alpha k} \leq 1,$$

and otherwise setting s to zero. This means that we keep in \tilde{f} all coefficients $c_{j,k,\psi}$ such that k is small enough (i.e., $\psi_{j,k}$ has low enough frequency) to satisfy (10), without regard to the relative sizes of the coefficients $c_{j,k,\psi}$. Since $B_2^\alpha(L_2(I)) = W^\alpha(L_2(I))$, this is an approximate solution of the nonparametric estimation problem [34].

The other interesting special case is when $1/p = \alpha/2 + 1/2$, so that $B_p^\alpha(L_p(I))$ has minimal smoothness to be embedded in $L_2(I)$. In this case, $\lambda = \mu$ and we set $s = t$ when $|t|^2 \geq \lambda |t|^p$, i.e.,

$$(11) \quad |t| \geq \lambda^{1/(2-p)},$$

and otherwise set s to zero. Here, we keep in \tilde{f} all coefficients $c_{j,k,\psi}$ above a certain threshold, without regard to the value of k , or equivalently, without regard to the frequency of $\psi_{j,k}$.

Motivated by the practical success of Rudin-Osher-Fatemi [33] in using $Y = \text{BV}(I)$ in (1), we set $\alpha = 1$ and find that $1/p = \alpha/2 + 1/2 = 1$, so we consider the space $Y = B_1^1(L_1(I))$. This space is very close to $\text{BV}(I)$, since

$$(12) \quad B_1^1(L_1(I)) \subset \text{BV}(I) \subset B_\infty^1(L_1(I)).$$

Thus, it is interesting to consider separately the case where $E(s) = |t - s|^2 + \mu |s|$ and $\lambda = \mu$. In this case, calculus shows that the exact minimizer of $E(s)$ is given by

$$s = \begin{cases} t - \lambda/2, & t > \lambda/2, \\ 0, & |t| \leq \lambda/2, \\ t + \lambda/2, & t < -\lambda/2. \end{cases}$$

Thus we shrink the wavelet coefficients $c_{j,k,\psi}$ toward zero by an amount $\lambda/2$ to obtain the exact minimizer $\tilde{c}_{j,k,\psi}$. This is precisely the *wavelet shrinkage* algorithm of Donoho and Johnstone [19]. Thus, wavelet shrinkage can be interpreted as the solution of the minimization problem (1) using $Y = B_1^1(L_1(I))$ with its wavelet-dependent norm.

In the spirit of searching for spaces of minimal smoothness for Y , we note that $B_1^1(L_1(I)) \subset B_q^1(L_1(I)) \subset L_2(I)$ when $1 \leq q \leq 2$, and not otherwise. (This can be easily derived from (4).) In fact, we can choose $Y = B_q^1(L_1(I))$ for any value of q ; if $q > 2$ then the spaces are not contained in $L_2(I)$, but if f is in $L_2(I)$ then any minimizer of (1) is necessarily in $L_2(I)$ because $\|f - g\|_{L_2(I)}$ is finite. Another issue of practical interest is whether the space Y contains *images with edges*, i.e., functions that are discontinuous across curves. $\text{BV}(I)$ does allow such images, but $B_q^1(L_1(I))$ does not for $q < \infty$.

When $Y = B_q^1(L_1(I))$, $1 < q < \infty$, substituting the equivalent norm (4) in (1) (with $s = 1$) yields

$$\sum_{j,k,\psi} |c_{j,k,\psi} - d_{j,k,\psi}|^2 + \lambda \left(\sum_k \left(\sum_{j,\psi} |d_{j,k,\psi}| \right)^q \right)^{1/q}.$$

This problem no longer decouples as it does for $Y = B_p^\alpha(L_p(I))$. On the other hand, calculus again shows that the minimizer $d_{j,k,\psi}$ satisfies

$$d_{j,k,\psi} = \begin{cases} 0, & |c_{j,k,\psi}| \leq \lambda_k \\ c_{j,k,\psi} - \lambda_k \operatorname{sgn}(c_{j,k,\psi}), & |c_{j,k,\psi}| > \lambda_k, \end{cases}$$

where $\operatorname{sgn}(x)$ is the usual sign function ($\operatorname{sgn}(x)$ is 1 when x is positive, -1 when x is negative, and zero when $x = 0$) and

$$\lambda_k := \frac{\lambda}{2} \left(\sum_{\ell} \left(\sum_{j,\psi} |d_{j,\ell,\psi}| \right)^q \right)^{-1+1/q} \left(\sum_{j,\psi} |d_{j,k,\psi}| \right)^{q-1}$$

is a *scale-dependent* shrinkage factor. If we denote by $q^* = q/(q-1)$ the dual exponent to q that satisfies $1/q + 1/q^* = 1$, then one finds that

$$\|(\lambda_k)\|_{\ell_{q^*}} := \left(\sum_k |\lambda_k|^{q^*} \right)^{1/q^*} = \frac{\lambda}{2}.$$

This result obviously holds when $q = 1$, for which we derived that $\lambda_k = \lambda/2$ for all k , so $\|(\lambda_k)\|_{\ell_\infty} = \sup_k |\lambda_k| = \lambda/2$.

As a practical matter, we explain how to solve two related variational problems. First, we consider

$$\|f - g\|_{L_2(I)}^2 + \lambda \|g\|_{B_2^1(L_1(I))}^2.$$

Using our usual substitution of wavelet coefficients, this reduces to setting

$$d_{j,k,\psi} = \begin{cases} 0, & |c_{j,k,\psi}| \leq \lambda_k \\ c_{j,k,\psi} - \lambda_k \operatorname{sgn}(c_{j,k,\psi}), & |c_{j,k,\psi}| > \lambda_k, \end{cases}$$

where

$$(13) \quad \lambda_k = \lambda \sum_{j,\psi} |d_{j,k,\psi}| = \lambda \sum_{\substack{j,\psi \\ |c_{j,k,\psi}| > \lambda_k}} (|c_{j,k,\psi}| - \lambda_k).$$

Thus, even though the problem does not decouple completely, as when $p = q$, we see that it decouples by scale.

Given λ and the set of coefficients $\{c_{j,k,\psi}\}_{j,\psi}$ at a given scale k , (13) is an implicit formula for λ_k that is easily solved. For example, one can sort $\{c_{j,k,\psi}\}_{j,\psi}$ in order of decreasing absolute value to obtain a sequence $\{a_j\}$ with $a_j \geq 0$; this takes $O(2^{2k} \log 2^{2k})$ operations at scale k . Next one examines each coefficient in turn; when it first happens that for a particular value $\mu = a_j$

$$\mu < \lambda \sum_{i=1}^j a_i - j\lambda\mu,$$

then one knows that λ_k is between a_j and the next larger coefficient, and can be found by solving a trivial linear equation. This takes at most $O(2^{2k})$ operations.

The second variational problem we examine is to minimize

$$(14) \quad \|f - g\|_{L_2(I)}^2 + \lambda \|g\|_{B_\infty^1(L_1(I))}.$$

Instead of solving this problem directly, we solve the following *dual problem*: find the minimum of

$$(15) \quad \|f - g\|_{L_2(I)}^2 \quad \text{given} \quad \|g\|_{B_\infty^1(L_1(I))} \leq M$$

for any fixed M . Any solution \tilde{f} of (14) is again a solution of (15) with $M = \|\tilde{f}\|_{B_\infty^1(L_1(I))}$. After applying the wavelet transform, this dual problem is to minimize

$$\sum_{j,k,\psi} |c_{j,k,\psi} - d_{j,k,\psi}|^2 \quad \text{given} \quad \sup_k \sum_{j,\psi} |d_{j,k,\psi}| \leq M.$$

Again the problem decouples by scale k . If, for a given k ,

$$\sum_{j,\psi} |c_{j,k,\psi}| \leq M,$$

then we obviously minimize at that scale by setting $d_{j,k,\psi} = c_{j,k,\psi}$ for all j and ψ . Otherwise, a continuity argument can be used to show that the minimizer at level k is also the minimizer of

$$\sum_{j,\psi} |c_{j,k,\psi} - d_{j,k,\psi}|^2 + \lambda_k \sum_{j,\psi} |d_{j,k,\psi}|$$

for some unknown λ_k ; we have already seen from our discussion of the $B_1^1(L_1(I))$ minimization problem that the solution is

$$d_{j,k,\psi} = \begin{cases} 0, & |c_{j,k,\psi}| \leq \lambda_k/2, \\ c_{j,k,\psi} - \lambda_k/2 \operatorname{sgn}(c_{j,k,\psi}), & |c_{j,k,\psi}| > \lambda_k/2. \end{cases}$$

We choose λ_k such that

$$\sum_{j,\psi} |d_{j,k,\psi}| = \sum_{\substack{j,\psi \\ |c_{j,k,\psi}| > \lambda_k/2}} (|c_{j,k,\psi}| - \frac{\lambda_k}{2}) = M;$$

an algorithm similar to that given for the $B_2^1(L_1(I))$ problem now suffices to find λ_k .

To each M we associate the value

$$\lambda = 2 \sum_k \lambda_k,$$

In Appendix I we show that λ is finite and that the solution g of our dual problem (15) is the minimizer of (14) with the associated value of λ .

4. Wavelet Representation of Images

The purpose of this section is to relate more directly wavelet-based image processing to the observed pixel values. Our view of a digitized image is that the pixel values (observations) are samples, which depend on the measuring device, of an intensity field $F(x)$ for x on the square $I = [0, 1]^2$. We start with the simplest model, that of a CCD camera, where the pixel samples are well modeled by averages of the intensity function F over small squares. Furthermore, let's consider, in this special case, the Haar wavelets on the square.

We assume that 2^{2m} pixel values p_j are indexed by $j = (j_1, j_2)$, $0 \leq j_1, j_2 < 2^m$ in the usual arrangement of 2^m rows and columns, and that each measurement is the average value of F on the subsquare covered by that

pixel. To fix notation, we note that the j th pixel covers the square $I_{j,m}$ with sidelength 2^{-m} and lower-left corner at the point $j/2^m$. We denote the characteristic function of I by $\chi := \chi_I$ and the $L_2(I)$ -normalized characteristic function of $I_{j,m}$ by $\chi_{j,m} := 2^m \chi_{I_{j,m}} = 2^m \chi(2^m \cdot - j)$.

We can write each pixel value as

$$\begin{aligned} p_j &= 2^{2m} \int \chi(2^m x - j) F(x) dx \\ &= 2^m \langle \chi_{j,m}, F \rangle. \end{aligned}$$

The standard practice in wavelet-based image processing is to use the observed pixel values p_j to create the function

$$\begin{aligned} f_m &= \sum_j p_j \chi(2^m \cdot - j) \\ &= \sum_j \langle \chi_{j,m}, F \rangle \chi_{j,m}, \end{aligned}$$

which we call the ‘‘observed image.’’ It follows that if the wavelet expansion of the intensity field F is

$$F = \sum_{0 \leq k} \sum_{j, \psi} c_{j,k,\psi} \psi_{j,k},$$

then the wavelet expansion of f is

$$f_m = \sum_{0 \leq k < m} \sum_{j, \psi} c_{j,k,\psi} \psi_{j,k}.$$

The main point is that f_m is the $L_2(I)$ projection of F onto $\text{span}\{\chi_{j,m}\}_j = \text{span}\{\psi_{j,k}\}_{0 \leq k < m, j, \psi}$. Furthermore, if F is in any function space whose norm is determined by a sequence norm of (Haar) wavelet coefficients, then so is f_m , and the sequence space norm of f_m is no greater than the sequence space norm of F .

One often uses wavelets that are smoother than the Haar wavelets. We consider this case now, without, however, changing our method of observation. In this more general setting, we assume for each scale k the existence of orthonormal scaling functions $\{\phi_{j,k}\}_j$ associated with the orthonormal wavelet basis $\{\psi_{j,k}\}_{j,k,\psi}$ of $L_2(I)$. Away from the boundary of I , $\phi_{j,k} = 2^k \phi(2^k \cdot - j)$ for a single ϕ with $\int \phi = 1$, and $\psi_{j,k} = 2^k \psi(2^k \cdot - j)$ for $\psi \in \Psi$; near the boundary we assume the existence of special boundary scaling functions and wavelets $\phi_{j,k}$ and $\psi_{j,k}$ as posited in Section 3. We also assume that $V_m := \text{span}\{\phi_{j,m}\}_j = \text{span}\{\psi_{j,k}\}_{0 \leq k < m, j, \psi}$ contains all polynomials degree $< r$ for some positive r . (For Haar wavelets, $r = 1$.) Thus,

$$F = \sum_{j,k,\psi} \langle F, \psi_{j,k} \rangle \psi_{j,k}.$$

For technical reasons we assume that there exists a constant C such that the support of each $\phi_{j,k}$ and $\psi_{j,k}$ is contained in a square with side-length $C2^{-k}$ that contains the point $j/2^k$, and that $\|\psi_{j,k}\|_{L_\infty(I)} \leq C2^k$ for all $\psi_{j,k}$; these conditions are satisfied for existing wavelet bases.

The $L_2(I)$ projection of F onto V_m is now

$$f_m = \sum_j \langle F, \phi_{j,m} \rangle \phi_{j,m} = \sum_{0 \leq k < m} \sum_{j, \psi} c_{j,k,\psi} \psi_{j,k}.$$

However, we have not measured $\langle F, \phi_{j,m} \rangle$ but $p_j = 2^m \langle F, \chi_{j,m} \rangle$, so our new ‘‘observed image’’ is

$$f_o = \sum_j \langle F, \chi_{j,m} \rangle \phi_{j,m},$$

which is *not* the $L_2(I)$ projection of F onto V_m . Thus, when we work with f_o rather than f_m , we would like to bound the error that is added by not using the correct $L_2(I)$ projection of the intensity field F .

The following principle was suggested by Cohen et al. [4]: If F is a polynomial of degree $< r$, then f_o should be a polynomial of degree $< r$. In [4] they presented an algorithm to modify the $\phi_{j,m}$ to satisfy this principle; here we show that this is sufficient to guarantee that the error in using f_o instead of f_m is bounded in a reasonable way.

To simplify our discussion, we consider the algorithm in one space dimension, i.e., $I = [0, 1]$. Extensions to two dimensions are straightforward but notationally cumbersome.

We first note that if P is a polynomial of any degree, then

$$p_\ell := \langle P, \chi_{\ell,m} \rangle = \bar{P}(\ell/2^m)$$

for a different polynomial \bar{P} of the same degree. Thus, if \bar{P} has degree $< r$, we would like

$$\sum_\ell p_\ell \phi_{\ell,m}(x)$$

to again be a polynomial of degree $< r$ for $x \in I$. This will be true away from the boundary of I , where $\phi_{j,m}(x) = 2^{m/2} \phi(2^m x - j)$, but not near the boundary of I . However, Cohen et al. point out that there exist invertible $r \times r$ matrices A_{left} and A_{right} such that when \bar{p}_ℓ is defined by

$$\begin{aligned} \begin{pmatrix} \bar{p}_0 \\ \vdots \\ \bar{p}_{r-1} \end{pmatrix} &:= A_{\text{left}} \begin{pmatrix} p_0 \\ \vdots \\ p_{r-1} \end{pmatrix}, \\ \begin{pmatrix} \bar{p}_{2^m-1} \\ \vdots \\ \bar{p}_{2^m-r} \end{pmatrix} &:= A_{\text{right}} \begin{pmatrix} p_{2^m-1} \\ \vdots \\ p_{2^m-r} \end{pmatrix}, \end{aligned}$$

and $\bar{p}_\ell = p_\ell$ otherwise, then

$$\sum_\ell \bar{p}_\ell \phi_{\ell,m}(x)$$

is again a polynomial of degree $< r$ for all $x \in I$.

This method of ‘‘preconditioning’’ the pixel values can be interpreted as changing the original orthogonal wavelet basis near the boundary to obtain new *biorthogonal* scaling functions $\phi_{j,k}$, dual scaling functions $\tilde{\phi}_{j,k}$, wavelets $\psi_{j,k}$, and dual wavelets $\tilde{\psi}_{j,k}$. Just as for the original wavelets $\psi_{j,k}$, we have for all polynomials P of degree $< r$,

$$(16) \quad \int_I P(x) \tilde{\psi}_{j,k}(x) dx = 0$$

We also have

$$f_m = \sum_j \langle F, \tilde{\phi}_{j,m} \rangle \phi_{j,m}$$

and

$$f_o = \sum_j \langle F, \chi_{j,m} \rangle \phi_{j,m}.$$

However,

$$\sum_{\ell} \langle P, \chi_{\ell,m} \rangle \phi_{\ell,m}$$

is a polynomial of degree $< r$ whenever P is. This algorithm for changing the orthogonal wavelets to biorthogonal wavelets can be extended to multiple dimensions by tensor products.

After constructing these biorthogonal wavelets, we consider the case of completely general measurements given by $\langle F, \tilde{\Phi}_{j,m} \rangle$ where $\tilde{\Phi}_{j,m}$ is the *point spread function* of the physical measuring device at location $j/2^m$; generally, this function has no relation whatsoever to $\tilde{\phi}_{j,m}$. We assume that the point spread functions $\tilde{\Phi}_{j,m}$ satisfy for constants C independent of j and m :

(1) The support of $\tilde{\Phi}_{j,m}$ contains the point $j/2^m$ and is contained in an interval of width $< C2^{-m}$.

(2) $\|\tilde{\Phi}_{j,m}\|_{L^\infty(I)} \leq C2^{m/2}$.

(3) Except for K values of $j/2^m$ near 0 and 1, $\tilde{\Phi}_{j,m}(x) = 2^{m/2}\tilde{\Phi}(2^m x - j)$ for a function $\tilde{\Phi}$, normalized so that $\int_{\mathbb{R}} \tilde{\Phi} = \int_{\mathbb{R}} \phi = 1$, and $2^m \geq 2K + 2r$

These conditions seem rather mild—they say that the point spread functions have bounded support, are not too big, are translation invariant away from the boundary, and are able to distinguish between polynomials of degree $< r$.

We note that for any polynomial P of degree $< r$, $p_\ell := \langle P, \tilde{\Phi}_{\ell,m} \rangle = \bar{P}(\ell/2^m)$ for another polynomial \bar{P} of the same degree, except for the K leftmost and K rightmost values of ℓ . This is because

$$\begin{aligned} \int_{\mathbb{R}} x^s \tilde{\Phi}(x - \ell) dx &= \int_{\mathbb{R}} (x + \ell)^s \tilde{\Phi}(x) dx \\ &= \sum_{i=0}^s \binom{s}{i} \ell^i \int_{\mathbb{R}} x^{s-i} \tilde{\Phi}(x) dx \end{aligned}$$

for all s and ℓ .

Because of our third assumption, the matrix

$$\begin{pmatrix} \langle 1, \tilde{\Phi}_{0,m} \rangle & \cdots & \langle 1, \tilde{\Phi}_{K+r-1,m} \rangle \\ \langle x, \tilde{\Phi}_{0,m} \rangle & \cdots & \langle x, \tilde{\Phi}_{K+r-1,m} \rangle \\ \vdots & \vdots & \vdots \\ \langle x^{r-1}, \tilde{\Phi}_{0,m} \rangle & \cdots & \langle x^{r-1}, \tilde{\Phi}_{K+r-1,m} \rangle \end{pmatrix}$$

has full rank, and so does the similar matrix at the right side of the interval. Thus, there exist invertible $(K+r) \times (K+r)$ matrices \bar{A}_{left} and \bar{A}_{right} such that if

$$\begin{pmatrix} \bar{p}_0 \\ \vdots \\ \bar{p}_{K+r-1} \end{pmatrix} = \bar{A}_{\text{left}} \begin{pmatrix} p_0 \\ \vdots \\ p_{K+r-1} \end{pmatrix},$$

$$\begin{pmatrix} \bar{p}_{2^m-1} \\ \vdots \\ \bar{p}_{2^m-K-r} \end{pmatrix} = \bar{A}_{\text{right}} \begin{pmatrix} p_{2^m-1} \\ \vdots \\ p_{2^m-K-r} \end{pmatrix},$$

and $\bar{p}_\ell = p_\ell$ otherwise, then $\bar{p}_\ell = \bar{P}(\ell/2^m)$ for all $\ell = 0, \dots, 2^m - 1$.

Applying these post-processing matrices \bar{A}_{left} and \bar{A}_{right} is equivalent to modifying $\tilde{\Phi}_{j,m}$ so that for any polynomial P of degree $< r$,

$$(17) \quad \int_I P \tilde{\Phi}_{\ell,m} = \bar{P}(\ell/2^m), \quad \ell = 0, \dots, 2^m - 1.$$

Numerically, we compute the inner products with the modified point spread functions by applying the appropriate post-processing matrices to the observed values near the boundary of I .

Thus, with these suitably modified point spread functions and scaling functions, we have for any polynomial P on I of degree $< r$,

$$\sum_{\ell} \langle P, \tilde{\Phi}_{j,m} \rangle \phi_{j,m}(x) = \bar{P}(x)$$

for a different polynomial \bar{P} on I of the same degree. We take our observed image to be

$$f_o = \sum_j \langle F, \tilde{\Phi}_{j,m} \rangle \phi_{j,m}.$$

In two dimensions, if the point spread function is of the form $\tilde{\Phi}_{j,m}(x, y) = \eta_{j_1,m}(x)\eta_{j_2,m}(y)$, $j = (j_1, j_2)$, then one can easily extend this construction using tensor products. In this case, one must apply pre- and post-conditioning to the pixel values in the $K+r$ rows and columns immediately adjacent to the boundaries of the image.

In Appendix II, we prove two things about f_o . The first is that f_o is close to f_m : If $\tilde{\Phi}_{j,m}$ and $\tilde{\phi}_{j,m}$ are *compatible to order s* , in the sense that

$$(18) \quad \int_I x^\gamma \tilde{\Phi}_{j,m} = \int_I x^\gamma \tilde{\phi}_{j,m} \quad 0 \leq j < 2^m, \quad 0 \leq |\gamma| < s,$$

then

$$(19) \quad \|f_o - f_m\|_{L_2(I)} \leq C2^{-m\alpha} \|F\|_{W^\alpha(L_2(I))}, \quad \alpha < s.$$

Since, by construction, $\int_I \tilde{\Phi}_{j,m} = \int_I \tilde{\phi}_{j,m}$ for all j , (18) is always true for $s = 1$, so (19) is true at least for $\alpha < 1$. Since the same result holds in two dimensions, and images with edges have $\alpha < 1/2$, this is sufficient for most purposes. If, in addition, away from the boundary of I , $\tilde{\Phi}_{j,m}$ and $\tilde{\phi}_{j,m}$ are symmetric about the point $(j + 1/2)/2^m$, then (18) holds for $s = 2$, and (19) holds for $\alpha < 2$. This condition is satisfied for many biorthogonal scaling functions (but not for Daubechies' orthogonal wavelets), and, we presume, for many point spread functions.

The second property is that f_o is just as smooth in Besov spaces as F . Specifically, for those values of $\alpha < r$ and $1/p < \alpha/d + 1$ in d dimensions such that $\|F\|_{B_p^\alpha(L_p(I))}$,

is equivalent to a sequence norm of the wavelet coefficients of F , then

$$(20) \quad \|f_o\|_{B_p^\alpha(L_p(I))} \leq C \|F\|_{B_p^\alpha(L_p(I))}.$$

With these two properties, all the theorems in this paper that are proved for f_m also hold for the ‘‘observed image’’ f_o , once one takes into account the difference between f_o and f_m bounded by (19).

5. Linear and Nonlinear Compression

In this section we briefly present some results from [12] (which are all classical, by this time; see [12] for references) that we need in the following sections.

The first result concerns what we call linear compression. We take for our approximation to F the wavelet approximation

$$f_K := \sum_{k < K} \sum_{j, \psi} c_{j,k,\psi} \psi_{j,k},$$

i.e., we include in the approximation all coefficients $c_{j,k,\psi}$ with frequency less than 2^K , $K \leq m$. This corresponds to progressive transmission, if one sends all coarse level wavelet coefficients before the finer level coefficients. We find that

$$\begin{aligned} \|F - f_K\|_{L_2(I)}^2 &= \sum_{k \geq K} \sum_{j, \psi} |c_{j,k,\psi}|^2 \\ &\leq \sum_{k \geq K} \sum_{j, \psi} \frac{2^{2\alpha k}}{2^{2\alpha K}} |c_{j,k,\psi}|^2 \\ &\leq 2^{-2\alpha K} \sum_k \sum_{j, \psi} 2^{2\alpha k} |c_{j,k,\psi}|^2 \\ &= 2^{-2\alpha K} \|F\|_{W^\alpha(L_2(I))}^2 \end{aligned}$$

Thus,

$$(21) \quad \|F - f_K\|_{L_2(I)} \leq 2^{-\alpha K} \|F\|_{W^\alpha(L_2(I))}$$

For our nonlinear compression algorithm, we take

$$f_\lambda := \sum_{\substack{k < m \\ |c_{j,k,\psi}| \geq \lambda}} c_{j,k,\psi} \psi_{j,k}.$$

Thus, we take all large coefficients, no matter their frequency, with the extra provision that we must work from the data, i.e., $k < m$. If we assume that $F \in B_q^\alpha(L_q(I))$, $1/q = \alpha/2 + 1/2$, then N , the number of coefficients greater than λ , satisfies

$$N\lambda^q \leq \sum_{j,k,\psi} |c_{j,k,\psi}|^q = \|F\|_{B_q^\alpha(L_q(I))}^q,$$

so

$$(22) \quad N \leq \lambda^{-q} \|F\|_{B_q^\alpha(L_q(I))}^q$$

and

$$\begin{aligned} \|f_\lambda - f_m\|_{L_2(I)}^2 &\leq \sum_{|c_{j,k,\psi}| < \lambda} |c_{j,k,\psi}|^2 \\ (23) \quad &\leq \sum_{|c_{j,k,\psi}| < \lambda} \lambda^{2-q} |c_{j,k,\psi}|^q \\ &\leq \lambda^{2\alpha/(\alpha+1)} \|F\|_{B_q^\alpha(L_q(I))}^q \end{aligned}$$

since $2 - q = 2\alpha/(1 + \alpha)$. If N is nonzero, then (22) implies

$$(24) \quad \lambda \leq N^{-1/q} \|F\|_{B_q^\alpha(L_q(I))}$$

and (23) and (24) yield

$$(25) \quad \|f_\lambda - f_m\|_{L_2(I)} \leq N^{-\alpha/2} \|F\|_{B_q^\alpha(L_q(I))}.$$

Note that, since there are 2^{2K} terms in f_K and N terms in f_λ , we get the same rate of approximation in each estimate; the only difference is that we use different norms to measure the smoothness of F . The nonlinear estimate is better for two reasons: for a given α , there are more images F in $B_q^\alpha(L_q(I))$ than there are in $B_2^\alpha(L_2(I)) = W^\alpha(L_2(I))$; and if a fixed image F is in $W^\alpha(L_2(I))$ then it's in $B_r^\beta(L_r(I))$, $1/r = \beta/2 + 1/2$, for some $\beta \geq \alpha$, i.e., a given image F will have greater smoothness in one of the nonlinear smoothness spaces $B_q^\alpha(L_q(I))$, $1/q = \alpha/2 + 1/2$ than in the Sobolev spaces $W^\alpha(L_2(I))$.

This analysis also applies to any compression scheme that satisfies

$$\tilde{f} = \sum \tilde{c}_{j,k,\psi} \psi_{j,k}$$

with

$$|c_{j,k,\psi} - \tilde{c}_{j,k,\psi}| \leq \lambda \text{ and } |c_{j,k,\psi}| < \lambda \implies \tilde{c}_{j,k,\psi} = 0.$$

Thus, it applies to threshold coding, zero-tree coding (see [11]), scalar quantization, and most types of vector quantization algorithms.

It is remarked in [6] that (25) is invertible, i.e., if we observe

$$\|f_\lambda - f_m\|_{L_2(I)} \leq C_1 N^{-\alpha/2}$$

for some α and C_1 , then one can conclude that $f_m \in B_q^\alpha(L_q(I))$, $1/q = \alpha/2 + 1/2$, and one can define an equivalent norm on $B_q^\alpha(L_q(I))$ such that in this norm $\|f_m\|_{B_q^\alpha(L_q(I))} = C_1$. (This statement is incorrect, but is close enough to the truth to be useful in practice; see [6] for the precise statement.) Thus, observing convergence rates for nonlinear wavelet approximations allows one to estimate the Besov smoothness of images.

Although the theory in [6] relates the image error to the number of nonzero coefficients, and not the number of bytes in the compressed file, the latter two quantities are closely related in practice; see, e.g., Figure 15 of [6]. The discussion there shows that, on average, only 6 bits were stored in the compressed image file for each nonzero coefficient. The computational results here provided a similar relationship between number of nonzero coefficients and file sizes of compressed images.

If F is also in $W^\beta(L_2(I))$ for some $\beta > 0$, then

$$\begin{aligned} \|F - f_\lambda\|_{L_2(I)}^2 &= \|f_m - f_\lambda\|_{L_2(I)}^2 + \|f_m - F\|_{L_2(I)}^2 \\ &\leq N^{-\alpha} \|F\|_{B_q^\alpha(L_q(I))}^2 + C2^{-2m\beta} \|F\|_{W^\beta(L_2(I))}^2. \end{aligned}$$

In other words, if

$$N^{-\alpha} \|F\|_{B_q^\alpha(L_q(I))}^2 \geq C2^{-2m\beta} \|F\|_{W^\beta(L_2(I))}^2$$

then the (unobserved) rate of convergence of $\|f_\lambda - F\|_{L_2(I)}$ is the same as the observed rate of convergence of $\|f_\lambda - f_m\|_{L_2(I)}$. Since $\beta \leq \alpha$, observing the rate of approximation for smaller N (middle bit rates) gives a better estimate for the smoothness of F than for the highest bit rates. By (6), if F is bounded, then $\beta \geq \alpha/(\alpha+1)$ and we get better estimates of convergence rates if

$$N \leq C2^{2m/(\alpha+1)} \|F\|_{B_q^\alpha(L_q(I))}^{2/\alpha} / \|F\|_{W^{\alpha/(\alpha+1)}(L_2(I))}^{2/\alpha}.$$

6. Error of Image Compression

In the preceding section, we gave a simple bound in $L_2(I)$ of the error in image compression using wavelets. Quantization strategies and error bounds are also available to minimize the error in $L_p(I)$, $0 < p < \infty$ (see [6]). The choice of error metric is a matter of much discussion; for example, in [6], it is argued that for when image quality of compressed natural scenes is judged by observers, $L_1(I)$ is a better metric and better matches the properties of the Human Visual System. For mammograms, however, it seems that optimizing the $L_2(I)$ metric leads to better diagnoses of compressed images, perhaps because of the wide spatial frequency range of objects of clinical interest; see preliminary results in [28].

Using $L_2(I)$ as the measure of image error is often criticized, since two images with equal $L_2(I)$ errors can be perceived as having vastly different visual quality, and images with different $L_2(I)$ errors can have similar visual quality.

We remark in this section that the Besov space norms $B_r^\beta(L_r(I))$, for $0 < \beta < \alpha$ and $1/r = \beta/2 + 1/2$, can also be used to measure the error between an original image f in $B_q^\alpha(L_q(I))$, $1/q = \alpha/2 + 1/2$, and a compressed image \tilde{f} . These spaces measure not only the $L_2(I)$ size of the error, but also the smoothness of the error. For example, assume that we have an original image f and two images, \tilde{f}_1 and \tilde{f}_2 , such that \tilde{f}_1 is derived from f by adding ϵ to each pixel, and \tilde{f}_2 is derived from f by adding Gaussian random noise to each pixel with mean 0 and variance ϵ^2 . Then

$$\|f - \tilde{f}_1\|_{L_2(I)}^2 = E(\|f - \tilde{f}_2\|_{L_2(I)}^2) = \epsilon^2,$$

yet $\|f - \tilde{f}_1\|_{B_r^\beta(L_r(I))} \approx \epsilon$ for all $\beta > 0$ while $E(\|f - \tilde{f}_2\|_{B_r^\beta(L_r(I))}) \rightarrow \infty$ as the number of pixels increases if $\beta > 0$.

Wavelet compression methods simultaneously give good approximation in all the spaces $B_r^\beta(L_r(I))$, $0 < \beta <$

α , $1/r = \beta/2 + 1/2$. Indeed, f_λ is the optimal approximation to f_m , in the sense that it minimizes the $B_r^\beta(L_r(I))$ norm of the error of any approximation with the same number of wavelet terms as f_λ . Furthermore, using the notation of the previous section, we see that

$$\begin{aligned} \|f_m - f_\lambda\|_{B_r^\beta(L_r(I))}^r &= \sum_{|c_{j,k,\psi}| < \lambda} |c_{j,k,\psi}|^r \\ &= \sum_{|c_{j,k,\psi}| < \lambda} |c_{j,k,\psi}|^{r-q} |c_{j,k,\psi}|^q \\ &\leq \lambda^{r-q} \|F\|_{B_q^\alpha(L_q(I))}^q \end{aligned}$$

Using (24), we have

$$\|f_m - f_\lambda\|_{B_r^\beta(L_r(I))}^r \leq N^{-(r-q)/q} \|F\|_{B_q^\alpha(L_q(I))}^r$$

so

$$\begin{aligned} \|f_m - f_\lambda\|_{B_r^\beta(L_r(I))} &\leq N^{-(1/q-1/r)} \|F\|_{B_q^\alpha(L_q(I))} \\ &= N^{-(\alpha-\beta)/2} \|F\|_{B_q^\alpha(L_q(I))}. \end{aligned}$$

This is an interesting idea: that a quantization strategy that is optimal for $L_2(I)$ gives good approximation in many smoothness spaces simultaneously. It may be possible that a quantization strategy developed for $L_2(I)$ yields good visual results not because it minimizes the $L_2(I)$ norm of the error, but because it minimizes the $B_r^\beta(L_r(I))$ norm of the error for some $0 < \beta < \alpha$. We note that the blocky error introduced by the JPEG standard at higher compression ratios has a large norm in $B_r^\beta(L_r(I))$ for $\beta > 0$.

7. Error, Smoothness, and Visual Perception

We believe that the perceived difference between an original image and a processed image is due not only to the size of the difference between the two images but also to the local changes in smoothness between images, and that such changes can be quantified in the Besov spaces $B_q^\alpha(L_q(I))$, $1/q = \alpha/2 + 1/2$. We discuss several examples: wavelet compression, pixel quantization, and the (artificial) example of adding one picture to another. In the first example, the error is most noticed in rough regions of the image (edges and textures), while in the second and third, the error is noticed more in smooth regions. This effect is usually explained as visual masking. We note that in all three examples, the error is noticed most where there is the largest change in the Besov smoothness of the image.

We first consider image compression by thresholding of wavelet coefficients. The compressed image is always smoother than the original ($\|f_\lambda\|_{B_q^\alpha(L_q(I))} \leq \|f\|_{B_q^\alpha(L_q(I))}$). Where the original image is smooth, the compressed image is about as smooth as the original, but along edges and in textured regions, the compressed image is much smoother than the original, and locally may have an order of smoothness $\beta \gg \alpha$. It is here, where the smoothness of the image changes (increases) the most, that the HVS perceives the most error.

In [6], DeVore, Jawerth, and Lucier considered the problem of the error induced by pixel quantization. In contrast to wavelet compression, pixel quantization usually *decreases* the smoothness of an image. DeVore et al. showed that if $f \in B_q^\alpha(L_q(I))$ and the measured image f_m has 2^{2m} pixels quantized to one of 2^n grey scale levels, then one needs $2^n \geq C2^{\alpha m}$ for the quantized image to have roughly the same smoothness as the original. Although stated for the image as a whole, this result holds in any subpart of the image. Thus, in smooth regions of the image, where α is large, one needs more grey scale levels to maintain the smoothness, while in rough regions one maintains smoothness with fewer quantization levels. Conversely, with the same number of quantization levels for all pixels in the image, quantization decreases the local smoothness more where the image is smooth than where it is rough. It is precisely where the smoothness changes (decreases) the most that the HVS perceives the most error, in the form of contouring.

As a third example, Rabbani considered the following problem in a talk. Consider one image f_1 (lenna, say) with all pixel values scaled between -20 and 20 , added, pixel by pixel, to another image f_2 (e.g., the mandrill). It is likely that, even locally, the combined image will have less smoothness than f_2 by itself. In this problem we are decreasing the smoothness of the image f_2 , and in the regions where f_2 is particularly smooth and f_1 is particularly rough, this decrease is most notable. And, yet again, it is precisely where the smoothness changes (decreases) the most that the HVS perceives the most difference between the combined image and the second image.

Our claim is more than that the HVS notices changes in smoothness in images—it is that it notices changes in smoothness as measured in the Besov spaces $B_q^\alpha(L_q(I))$, $1/q = \alpha/2 + 1/2$, or more generally, $1/q = \alpha/2 + 1/p$ for some $0 < p \leq \infty$. One could define perfectly good (if somewhat strange) smoothness norms based on the size of the coefficients of the block DCT transforms used in JPEG image compression (see, e.g., [15]), and it would happen that images compressed with the JPEG algorithm will always be smoother than the original image *in these smoothness spaces*. However, the HVS does not always perceive the images as smoother, since it notices the blocky artifacts at high compression ratios. Thus, one might say that implied smoothness model built into wavelet image compression matches the perceptual properties of the HVS better than the implied smoothness model of the JPEG algorithm.

8. Linear and Nonlinear Noise Removal

In the previous sections, we assumed that the measured pixel values p_j were the exact values of $2^m \langle F, \phi_{j,m} \rangle$ (averages over squares in the Haar case). In this section, we assume that our measurements are corrupted by Gaussian noise, that is, that we measure not p_j , but $\bar{p}_j := p_j + \eta_j$, where ϵ_j are i.i.d. normal random variables with mean 0

and variance σ_0^2 (denoted by $N(0, \sigma_0^2)$). From this we construct

$$\bar{f} = \sum_j [\langle F, \phi_{j,m} \rangle + 2^{-m} \eta_j] \phi_{j,m}.$$

Since the wavelet transform that takes $\{\langle F, \phi_{j,m} \rangle\} \rightarrow \{c_{j,k,\psi}\}_{0 \leq k < m, j, \psi}$ is an orthonormal transformation, we have that

$$\bar{f} = \sum_{0 \leq k < m} \sum_{j, \psi} \bar{c}_{j,k,\psi} \psi_{j,k}$$

and $\bar{c}_{j,k,\psi} = c_{j,k,\psi} + \epsilon_{j,k,\psi}$, where the $\epsilon_{j,k,\psi}$ are i.i.d. $N(0, \sigma_0^2/2^{2m})$ random variables. This model assumes that the expected value of the noise,

$$E(\|f_m - \bar{f}\|_{L_2(I)}^2) = \sigma_0^2,$$

is independent of the number of pixels. We now examine how the linear and nonlinear algorithms can be applied to \bar{f} to achieve noise removal.

Starting with \bar{f} , the linear algorithm calculates

$$\tilde{f} = \sum_{k < K} \bar{c}_{j,k,\psi} \psi_{j,k};$$

we choose K to minimize $E(\|f_m - \tilde{f}\|_{L_2(I)}^2)$. Using the wavelet decompositions of \tilde{f} and f_m , we calculate

$$\begin{aligned} E(\|f_m - \tilde{f}\|_{L_2(I)}^2) &= \sum_{k < K} \sum_{j, \psi} E([c_{j,k,\psi} - \bar{c}_{j,k,\psi}]^2) \\ &\quad + \sum_{K \leq k < m} \sum_{j, \psi} |c_{j,k,\psi}|^2 \\ &\leq \sum_{k < K} \sum_{j, \psi} E(\epsilon_{j,k,\psi}^2) \\ &\quad + 2^{-2\alpha K} \|F\|_{W^\alpha(L_2(I))}^2 \\ &= 2^{2K-2m} \sigma_0^2 + 2^{-2\alpha K} \|F\|_{W^\alpha(L_2(I))}^2. \end{aligned}$$

The inequality follows from (21), and the second equality holds because the 2^{2K} random variables $\epsilon_{j,k,\psi}$ each have variance $2^{-2m} \sigma_0^2$.

We set $N := 2^{2K}$ and we minimize $E(\|f_m - \tilde{f}\|_{L_2(I)}^2)$ with respect to N . Calculus shows that we overestimate the error at most by a factor of 2 if we set the two terms in our bound equal to each other, i.e., $N \sigma_0^2 2^{-2m} = N^{-\alpha} \|F\|_{W^\alpha(L_2(I))}^2$. This yields

$$N = \left(\frac{\|F\|_{W^\alpha(L_2(I))}^2 2^{2m}}{\sigma_0^2} \right)^{1/(\alpha+1)},$$

and

$$E(\|f_m - \tilde{f}\|_{L_2(I)}^2) \leq 2(2^{-2m} \sigma_0^2)^{\alpha/(\alpha+1)} \|F\|_{W^\alpha(L_2(I))}^{2/(\alpha+1)}.$$

Using (21) again with $K = m$ gives

$$\begin{aligned} E(\|F - \tilde{f}\|_{L_2(I)}^2) &\leq 2(2^{-2m} \sigma_0^2)^{\alpha/(\alpha+1)} \|F\|_{W^\alpha(L_2(I))}^{2/(\alpha+1)} \\ &\quad + 2^{-2\alpha m} \|F\|_{W^\alpha(L_2(I))}^2. \end{aligned}$$

This linear algorithm removes all terms $c_{j,k,\psi} \psi_{j,k}$ with k greater than or equal to a threshold K ; these terms can

be considered to have frequency at least 2^K . Thus, the linear method considers any low-frequency structure to be signal, and any high-frequency structure to be noise, no matter how large $c_{j,k,\psi}$, the scaled amplitude of the signal, might be. This is not acceptable to people, such as astronomers, who deal with high amplitude, small extent (and hence high frequency) signals. Some astronomical researchers have proposed eliminating the low-order bit planes to achieve noise removal and compression if these bit planes have entropy close to one. This will remove all low amplitude features, no matter how great their extent, for example, variations in the level of background radiation. The nonlinear algorithm presented next, which employs Donoho and Johnstone's wavelet shrinkage, recognizes both high-amplitude, high-frequency structures and low-amplitude, low-frequency structures as signals. Similar algorithms have been used in astronomical calculations, e.g., by White [35].

We define the shrinkage operator

$$s_\lambda(t) := \begin{cases} t - \lambda, & t > \lambda, \\ 0, & |t| \leq \lambda, \\ t + \lambda, & t < -\lambda. \end{cases}$$

Our noise-removed image is

$$\begin{aligned} \tilde{f}_m &= \sum_{k < m} \sum_{j, \psi} s_\lambda(\bar{c}_{j,k,\psi}) \psi_{j,k} \\ &= \sum_{k < m} \sum_{j, \psi} s_\lambda(c_{j,k,\psi} + \epsilon_{j,k,\psi}) \psi_{j,k} \end{aligned}$$

where λ is to be determined, and the error is

$$\|f_m - \tilde{f}_m\|_{L_2(I)}^2 = \sum_{k < m} \sum_{j, \psi} |c_{j,k,\psi} - s_\lambda(c_{j,k,\psi} + \epsilon_{j,k,\psi})|^2.$$

Since for all t ,

$$|t - s_\lambda(t + \epsilon)| = \begin{cases} |\epsilon - \lambda|, & \lambda - t \leq \epsilon, \\ |t|, & -\lambda - t \leq \epsilon \leq \lambda - t, \\ |\epsilon + \lambda|, & \epsilon \leq -\lambda - t, \end{cases}$$

one has

$$(26) \quad |t - s_\lambda(t + \epsilon)|^2 \leq \begin{cases} (\epsilon - \lambda)^2, & t > \lambda, \\ \max(t^2, s_\lambda^2(\epsilon)), & |t| \leq \lambda, \\ (\epsilon + \lambda)^2, & t < -\lambda. \end{cases}$$

Thus, if ϵ is normally distributed with mean zero and variance σ^2 , with probability distribution $P_\sigma(\epsilon)$,

$$(27) \quad E(|t - s_\lambda(t + \epsilon)|^2) \leq \begin{cases} \lambda^2 + \sigma^2, & |t| > \lambda, \\ t^2 + E(s_\lambda^2(\epsilon)), & |t| \leq \lambda. \end{cases}$$

In estimating the error of noise removal by wavelet shrinkage, we now apply (27) to the case where $t = c_{j,k,\psi}$, $\epsilon = \epsilon_{j,k,\psi}$, $\sigma^2 = \sigma_0^2/2^{2m}$, and $\lambda = a\sigma_0/2^m$, where a is to be determined.

Recall that if N denotes the number of coefficients $c_{j,k,\psi}$ with $|c_{j,k,\psi}| \geq \lambda$,

$$(28) \quad N \leq \lambda^{-q} \|F\|_{B_q^\alpha(L_q(I))}^q,$$

while

$$(29) \quad \sum_{|c_{j,k,\psi}| \leq \lambda} |c_{j,k,\psi}|^2 \leq \lambda^{2-q} \|F\|_{B_q^\alpha(L_q(I))}^q$$

Combining (27), (28), and (29) yields

$$\begin{aligned} (30) \quad E(\|f_m - \tilde{f}_m\|_{L_2(I)}^2) &\leq \sum_{|c_{j,k,\psi}| > \lambda} (\lambda^2 + \sigma^2) \\ &\quad + \sum_{|c_{j,k,\psi}| \leq \lambda} [c_{j,k,\psi}^2 + E(s_\lambda^2(\epsilon_{j,k,\psi}))] \\ &\leq \lambda^{-q} \|F\|_{B_q^\alpha(L_q(I))}^q (\lambda^2 + \sigma^2) \\ &\quad + \lambda^{2-q} \|F\|_{B_q^\alpha(L_q(I))}^q \\ &\quad + 2^{2m} 2 \int_\lambda^\infty (x - \lambda)^2 P_\sigma(x) dx \\ &= \frac{\sigma_0^{2-q}}{2^{m(2-q)}} \|F\|_{B_q^\alpha(L_q(I))}^q [2a^{2-q} + a^{-q}] \\ &\quad + \sigma_0^2 2 \int_a^\infty (x - a)^2 P_1(x) dx, \end{aligned}$$

where we have bounded the number of coefficients with $|c_{j,k,\psi}| \leq \lambda$ simply by 2^{2m} .

Inequality (30) is our main estimate. We note, and emphasize, that given only two parameters characterizing the smoothness of an image (α , from which we derive $q = 2/(\alpha + 1)$), and $\|F\|_{B_q^\alpha(L_q(I))}$ and an estimate of the standard deviation of the noise in the image, one can numerically minimize (30) with respect to a and use $a\sigma_0/2^m$ as the value of λ that minimizes our bound on the error. We apply this technique in Section 10 to various images.

Using the symbolic manipulation package Maple, we find that the last term in (30) is bounded by

$$(31) \quad \sigma_0^2 \frac{4}{a^3} P_1(a) = \sigma_0^2 \frac{4}{a^3} \frac{1}{\sqrt{2\pi}} e^{-a^2/2}$$

for all $a > 1$; in fact (31) is the first term of the asymptotic expansion of

$$\sigma_0^2 2 \int_a^\infty (x - a)^2 P_1(x) dx,$$

and

$$\sigma_0^2 \frac{4}{a^3} \frac{1}{\sqrt{2\pi}} e^{-a^2/2} - \sigma_0^2 2 \int_a^\infty (x - a)^2 P_1(x) dx = O(1/a^5)$$

as $a \rightarrow \infty$.

We can get a simple approximation to the critical a and a bound on the error. One can determine a so that

$$\frac{\sigma_0^{2-q}}{2^{m(2-q)}} \|F\|_{B_q^\alpha(L_q(I))}^q = \sigma_0^2 e^{-a^2/2}$$

or

$$(32) \quad a = \sqrt{(2-q) \ln 2^{2m} - 2q \ln \frac{\|F\|_{B_q^\alpha(L_q(I))}}{\sigma_0}}.$$

With this a , we have

$$E(\|f_m - \tilde{f}_m\|_{L_2(I)}^2) \leq \frac{\sigma_0^{2-q}}{2^{m(2-q)}} \|F\|_{B_q^\alpha(L_q(I))}^q [2a^{2-q} + a^{-q} + 4/\sqrt{2\pi}a^{-3}].$$

If we assume that $\sigma_0 \leq \|F\|_{B_q^\alpha(L_q(I))}$ and 2^m is large enough that

$$(4/\sqrt{2\pi})^{1/3} \approx 1.17 \leq a \leq \sqrt{(2-q) \ln 2^{2m}}$$

then

$$E(\|f_m - \tilde{f}_m\|_{L_2(I)}^2) \leq 2 \left(\frac{\sigma_0^2}{2^{2m}} \right)^{\alpha/(\alpha+1)} \|F\|_{B_q^\alpha(L_q(I))}^{2/(\alpha+1)} \left[\left(\frac{2\alpha}{\alpha+1} \ln 2^{2m} \right)^{\alpha/(\alpha+1)} + 1 \right],$$

since $q = 2/(\alpha+1)$ and $2-q = 2\alpha/(\alpha+1)$. If F is bounded, then (6) and (21) show that

$$\begin{aligned} E(\|F - \tilde{f}_m\|_{L_2(I)}^2) &= \|F - f_m\|_{L_2(I)}^2 + E(\|f_m - \tilde{f}_m\|_{L_2(I)}^2) \\ &\leq C 2^{-2m\alpha/(\alpha+1)} \|F\|_{L_\infty(I)}^{2\alpha/(\alpha+1)} \|F\|_{B_q^\alpha(L_q(I))}^{2/(\alpha+1)} \\ &\quad + E(\|f_m - \tilde{f}_m\|_{L_2(I)}^2). \end{aligned}$$

Thus, we achieve the same rate of approximation to the real intensity field F as we do to the sampled image f_m .

In equation (32) the quantity

$$(33) \quad \frac{\|F\|_{B_q^\alpha(L_q(I))}}{\sigma_0}$$

arises quite naturally. Insofar as α can be interpreted as a measure of the *structure* in an image, $\|F\|_{B_q^\alpha(L_q(I))}$ can be interpreted as the amount of *information* in an image. Thus, we interpret (33) as a pertinent (and important) new signal-to-noise ratio that quantifies the visual effects of adding noise to an image more reliably than the usual signal-to-noise ratio based on the $L_2(I)$ norm of F . This quantity also arises naturally in an analysis of the wavelet-vaguelette transform together with wavelet shrinkage applied to homogeneous integral equations—see [27].

We remark that a similar analysis can be applied to the wavelet truncation method of noise removal proposed in [12]. In this case,

$$\tilde{f}_m = \sum_{j,k,\psi} t_\lambda(c_{j,k,\psi} + \epsilon_{j,k,\psi})$$

and

$$\|f_m - \tilde{f}_m\|_{L_2(I)}^2 = \sum_{j,k,\psi} |c_{j,k,\psi} - t_\lambda(c_{j,k,\psi} + \epsilon_{j,k,\psi})|^2,$$

where t_λ is the truncation function

$$t_\lambda(s) = \begin{cases} s, & |s| > \lambda, \\ 0, & |s| \leq \lambda. \end{cases}$$

We have

$$|t_\lambda(s + \epsilon) - s| = \begin{cases} |\epsilon|, & |s + \epsilon| > \lambda, \\ |s|, & |s + \epsilon| \leq \lambda. \end{cases}$$

A rather crude inequality is

$$|t_\lambda(s + \epsilon) - s| \leq \begin{cases} \max(|s|, |t_{\lambda/2}(\epsilon)|), & |s| \leq 2\lambda, \\ 2|\epsilon|, & |s| > 2\lambda, \end{cases}$$

and if ϵ is normally distributed with mean zero and variance σ^2 ,

$$E(|s - t_\lambda(s + \epsilon)|^2) \leq \begin{cases} s^2 + E(t_{\lambda/2}^2(\epsilon)), & |s| \leq 2\lambda, \\ 4\sigma^2, & |s| > 2\lambda. \end{cases}$$

By following the rest of our analysis for wavelet shrinkage, the reader can discover new choices of λ and new error bounds, which have the same rate of convergence as our bounds for wavelet shrinkage.

Since we have 2^{2m} observations, Donoho and Johnstone have suggested using

$$a = \sqrt{2 \ln 2^{2m}}$$

as a “universal” choice for a in their VisuShrink method. These two suggestions agree as $q = 2/(\alpha+1) \rightarrow 0$, i.e., as $\alpha \rightarrow \infty$. This would seem to be a good choice for the examples given in [19] and [21], as their first three sample signals (Blocks, Bumps, and HeaviSine) are in the one-dimensional Besov spaces $B_q^\alpha(L_q)$, $1/q = \alpha + 1/2$, for all $\alpha > 0$. That is, in spite of the discontinuities and peaks in these sample functions, they are *infinitely smooth* in the scale of spaces $B_q^\alpha(L_q)$. However, images with edges have severe inherent limitations in smoothness, since $\alpha < 1$ [6], so $2\alpha/(\alpha+1) < 1$ and the smoothing parameter in [19] [21] results in over-smoothing. In fact, our estimates of the smoothness of images in [6] and several examples here suggest that for many images $.3 \leq \alpha \leq .7$, so the smoothing parameter should be even smaller. At high signal-to-noise ratios, with $\|F\|_{B_q^\alpha(L_q(I))}/\sigma_0 \gg 1$, the smoothing parameter should be reduced even more.

Conversely, for very smooth data, or very high noise levels, $\|F\|_{B_q^\alpha(L_q(I))}/\sigma_0 \ll 1$, and Donoho’s suggestion does not smooth enough to recover the smooth function F . See, however, the arguments in [16], where it is shown that with high probability, the universal choice suggested above leads to reconstructed functions with the same smoothness as F .

If we ignore the change in λ due to the signal-to-noise ratio, then our error bound with

$$a = \sqrt{\frac{2\alpha}{\alpha+1} \ln 2^{2m}}$$

is smaller than the bound achievable with $a = \sqrt{2 \ln 2^{2m}}$ only by a factor of x^x , where $0 < x = \alpha/(\alpha+1) < 1$. Since x^x achieves its minimum of about .69 when $x = 1/e$, the error bounds are not all that different. However, the greater error using $a = \sqrt{2 \ln 2^{2m}}$ is introduced by shrinking the real image coefficients more than necessary, so the effect is quite noticeable visually.

9. Noise Removal and Variational Problems with Biorthogonal Wavelets

Biorthogonal wavelets are quite attractive for image processing for several reasons. First, although symmetric orthogonal wavelets do not exist, symmetric biorthogonal wavelets can be constructed, and symmetry seems to be a property that reduces the visual perception of errors in processed images. Second, some biorthogonal wavelets lend themselves to very fast fixed-point arithmetic algorithms for their computation. Third, one can sometimes achieve very high approximation order with biorthogonal wavelets that oscillate very little, in contrast to the Daubechies' orthogonal wavelets, which oscillate more and more as the approximation order increases.

For these reasons, we would like to extend the theory in the preceding sections to biorthogonal wavelets. We first consider (real) biorthogonal wavelets on \mathbb{R} as defined, e.g., by Cohen, Daubechies, and Feauveau [3] and Herley and Vetterli [25]. There is a wavelet ψ , a dual wavelet $\tilde{\psi}$, a scaling function ϕ and a dual scaling function $\tilde{\phi}$. As before, we define $\psi_{j,k}(x) = 2^{k/2}\psi(2^kx - j)$; similarly for $\tilde{\psi}_{j,k}$, $\phi_{j,k}$, and $\tilde{\phi}_{j,k}$. Any function f in $L_2(\mathbb{R})$ can be written as

$$f = \sum_{j,k \in \mathbb{Z}} \langle f, \tilde{\psi}_{j,k} \rangle \psi_{j,k} \quad \text{and}$$

$$f = \sum_{j \in \mathbb{Z}, k \geq 0} \langle f, \tilde{\psi}_{j,k} \rangle \psi_{j,k} + \sum_{j \in \mathbb{Z}} \langle f, \tilde{\phi}_{j,0} \rangle \phi_{j,0}.$$

One chooses a wavelet–dual wavelet pair such that ψ has high smoothness properties, and $\tilde{\psi}$ has high numbers of zero moments. It can be shown that

$$(34) \quad \|f\|_{L_2(\mathbb{R})}^2 \asymp \sum_{j,k \in \mathbb{Z}} |\langle f, \tilde{\psi}_{j,k} \rangle|^2 \quad \text{and}$$

$$\|f\|_{L_2(\mathbb{R})}^2 \asymp \sum_{j \in \mathbb{Z}, k \geq 0} |\langle f, \tilde{\psi}_{j,k} \rangle|^2 + \sum_{j \in \mathbb{Z}} |\langle f, \tilde{\phi}_{j,0} \rangle|^2.$$

The set $\{\psi_{j,k}\}_{j,k \in \mathbb{Z}}$ now forms a Riesz basis for $L_2(\mathbb{R})$, rather than an orthonormal basis. For orthogonal wavelets, $\psi = \tilde{\psi}$, $\phi = \tilde{\phi}$, and we have equality in (34).

One constructs two-dimensional biorthogonal wavelets in the same way as orthogonal wavelets, by tensor products. And, again, it is possible to apply ideas similar to those found in [4] to modify $\psi_{j,k}$ and $\tilde{\psi}_{j,k}$ and construct biorthogonal wavelets on the square I with nice properties; see, e.g., [26]. One can show that there exist positive constants A and B such that for every f in $L_2(I)$,

$$(35) \quad f = \sum_{j,k,\psi} c_{j,k,\psi} \psi_{j,k} \quad \text{and}$$

$$A \left(\sum_{j,k,\psi} c_{j,k,\psi}^2 \right)^{1/2} \leq \|f\|_{L_2(I)} \leq B \left(\sum_{j,k,\psi} c_{j,k,\psi}^2 \right)^{1/2},$$

where $c_{j,k,\psi} = \int_I f \tilde{\psi}_{j,k}$.

Again, one can determine whether a function f is in the Besov space $B_q^\alpha(L_p(I))$ by examining the biorthogonal wavelet coefficients of f . In particular, there is a number r

that depends on ψ , $\tilde{\psi}$, ϕ , and $\tilde{\phi}$ such that if $2/p - 2 < \alpha < r$ (so that $B_q^\alpha(L_p(I))$ is embedded in $L_s(I)$ for some $s > 1$), then

$$(36) \quad \|f\|_{B_q^\alpha(L_p(I))} \asymp \left(\sum_k \left(\sum_{j,\psi} 2^{k(\alpha p + p - 2)} |c_{j,k,\psi}|^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}}.$$

When $p = q$,

$$(37) \quad \|f\|_{B_p^\alpha(L_p(I))} \asymp \left(\sum_k \sum_{j,\psi} 2^{k(\alpha p + p - 2)} |c_{j,k,\psi}|^p \right)^{1/p}.$$

Again, the expression on the right of (5) is wavelet-dependent.

One notes that when $p = q = 2$, then

$$\|f\|_{B_2^\alpha(L_2(I))} \asymp \left(\sum_k \sum_{j,\psi} 2^{2\alpha k} |c_{j,k,\psi}|^2 \right)^{1/2},$$

and when $p = q$ and $1/p = \alpha/2 + 1/2$, then $\alpha p + p - 2 = 0$ and

$$\|f\|_{B_q^\alpha(L_q(I))} \asymp \left(\sum_k \sum_{j,\psi} |c_{j,k,\psi}|^q \right)^{1/q}.$$

We now examine how using biorthogonal wavelets affects the analysis of the preceding sections.

In Section 3 on variational problems, we replaced the Besov space norms with equivalent sequence norms; because we used orthogonal wavelets, the $L_2(I)$ norms were equal to the ℓ_2 sequence norms of the coefficients. We obtain equivalent problems if we replace the $L_2(I)$ norms with the equivalent ℓ_2 norms of the *biorthogonal* wavelet coefficients, so all the calculations in that section apply as well to biorthogonal wavelets.

For the same reason, the analysis of compression schemes in Section 5 goes through without difficulty. The only difference is in the final estimates; for example, (25) would appear as

$$\|f_\lambda - f_m\|_{L_2(I)} \leq B N^{-\alpha/2} \|F\|_{B_q^\alpha(L_q(I))},$$

where B is the $L_2(I)$ -norm equivalence constant given in (35).

The analysis of biorthogonal wavelets in noise-removal computations is slightly more interesting. Here, we assume that the underlying image intensity field F is expressed as

$$F = \sum_{j,k,\psi} \langle F, \tilde{\psi}_{j,k} \rangle \psi_{j,k}$$

and that an observed (noise-free) image is

$$f = \sum_{j,\psi} \langle F, \tilde{\phi}_{j,m} \rangle \phi_{j,m}.$$

If we assume that the observed pixel values are corrupted by i.i.d. Gaussian noise $\eta_{j,m,\psi}$ with mean zero and variance σ_0^2 , then the observed noisy image is

$$f = \sum_{j,\psi} [\langle F, \tilde{\phi}_{j,m} \rangle + 2^{-m} \eta_{j,m,\psi}] \phi_{j,m}.$$

The transformation that takes $\{\langle F, \tilde{\phi}_{j,m} \rangle + 2^{-m}\eta_{j,m,\psi}\}_{j,\psi}$ to $\{\langle F, \tilde{\psi}_{j,k} \rangle + \epsilon_{j,k,\psi}\}_{j,k < m,\psi}$ is no longer orthonormal, so while the random variables $\epsilon_{j,k,\psi}$ are Gaussian (since the transformation is linear), they are neither independent nor identically distributed.

By applying wavelet shrinkage to the noisy coefficients of f , we obtain the noise-reduced image

$$f_\lambda = \sum_{j,k,\psi} s_\lambda(c_{j,k,\psi} + \epsilon_{j,k,\psi})\psi_{j,k}$$

and the square of the $L_2(I)$ error is

$$\|f - f_\lambda\|_{L_2(I)}^2 \leq B^2 \sum_{j,k,\psi} |c_{j,k,\psi} - s_\lambda(c_{j,k,\psi} + \epsilon_{j,k,\psi})|^2.$$

Thus we need to bound

$$\begin{aligned} E\left(\sum_{j,k,\psi} |c_{j,k,\psi} - s_\lambda(c_{j,k,\psi} + \epsilon_{j,k,\psi})|^2\right) \\ = \sum_{j,k,\psi} E(|c_{j,k,\psi} - s_\lambda(c_{j,k,\psi} + \epsilon_{j,k,\psi})|^2) \end{aligned}$$

where both expectations are taken with respect to the joint probability distribution $P(\{\epsilon_{j,k,\psi}\}_{j,k,\psi})$.

Each expectation is applied to a single term that depends only on $\epsilon_{j,k,\psi}$. Thus, we can integrate out the dependence on the other random variables and show that the preceding expression is equal to

$$\sum_{j,k,\psi} E_{\epsilon_{j,k,\psi}}(|c_{j,k,\psi} - s_\lambda(c_{j,k,\psi} + \epsilon_{j,k,\psi})|^2),$$

where $E_{\epsilon_{j,k,\psi}}$ is the expectation with respect to the marginal distribution of $\epsilon_{j,k,\psi}$, which is normal with mean zero.

Now, (27) still holds, and, as pointed out by Donoho in [17], the stability condition (35) (which Donoho calls the near-orthonormality property) allows one to show that there exist positive constants C_1 and C_2 such that

$$C_1 \frac{\sigma_0^2}{2^{2m}} \leq \text{Var}(\epsilon_{j,k,\psi}) \leq C_2 \frac{\sigma_0^2}{2^{2m}}$$

for all ψ , j , and k . Thus, one can follow the line of argument in Section 8 after (27) to arrive at the same asymptotic rate of convergence of the noise removal algorithms with different constants in the final bounds.

10. Noise Removal Computations

We conducted experiments using wavelet shrinkage to remove Gaussian noise from some images. Our main conclusion is that shrinkage parameters chosen by minimizing (30) lead to less shrinkage (smoothing), smaller errors, and better images than using the parameter suggested by Donoho and Johnstone in *VisuShrink* or than given by (32).

Our main computations are applied to the 24 images on the Kodak *Photo CD Photo Sampler*, Final Version 2.0, widely distributed by Apple Computer Corporation with its Macintosh computers. We propose that researchers in image processing consider these images as test images for

new algorithms. The images are rather large (2048 by 3072 pixels), are of relatively high quality, cover a range of subjects (although all are of natural scenes), and, as we shall see, are of varied smoothness. All images on the CD have been released for any image processing use. It's not clear how the lossy compression algorithms applied to images in sizes above 512 by 768 pixels has affected the images for test purposes; we still believe they are of significantly higher quality than, e.g., *lenna*.

We used the program `hpcdtoppm -ycc` to extract the intensity component of each Photo CD image **img0001** to **img0024** at the 2048 by 3072 size. The intensity images were not gamma-corrected before being used in our tests. Smaller images were calculated by averaging pixel values over 2×2 , 4×4 , ... squares of pixels.

We use the fifth-order-accurate 2–10 biorthogonal wavelets illustrated on page 272 of [5] (cf. [3] and [25]). These wavelets have several theoretical and practical advantages for image processing. First, if we assume that measured pixel values are true averages of an underlying intensity field F , then the wavelet coefficients we calculate are exactly those of F , since the dual functionals of these wavelets are piecewise constant. Although not orthogonal, these wavelets are not too far from orthogonal; they have few oscillations; and they decay rapidly away from zero, looking like smoothed versions of the Haar wavelet. They lend themselves to fast fixed-point computations, since the wavelet coefficients are dyadic rationals with small numerators. We modified the wavelets at the boundary in a way equivalent to reflecting the image across each side, so that our scheme is formally first-order accurate at boundaries.

We first estimated the smoothness of each image as discussed in Section 5. After instrumenting our compression program to report the number of nonzero coefficients N in each compressed image f_N , we compressed each image at various compression levels (from about 150 to 1,000,000 nonzero coefficients) using scalar quantization of the wavelet coefficients. We then calculated the best least-squares line that expressed the relationship between error and the number of nonzero coefficients on a log-log scale, i.e.,

$$\|f - f_N\|_{L_2(I)} \approx CN^{-\alpha/2}.$$

We estimated that F (the true intensity field) is in the Besov space $B_q^\alpha(L_q(I))$, $1/q = \alpha/2 + 1/2$, and that $\|F\|_{B_q^\alpha(L_q(I))} = C$. These values are reported in Table 1 for each image; we report the correlation coefficient (rounded to the nearest hundredth) as a guide to how well the error vs. number-of-nonzero-coefficients curve was approximated by a straight line.

Since our noise removal error estimates depend on the resolution at which we sample the image, we downsampled each 2048×3072 image by averaging to obtain 1024×1536 , 512×768 , 256×384 , 128×192 , and 64×96 images. Our rectangular images do not satisfy our previous assumption that images are square with 2^m pixels on a side; nonetheless, we applied our formulas after substituting M , the number of pixels in our images, for 2^{2m} .

TABLE 1
Shrinkage Parameters and Errors

λ_V	E_V	λ_e	E_e	λ_c	E_c	$\frac{ \lambda_e - \lambda_c }{\lambda_c} \leq .1$
img0001: $\alpha = 0.5536$, $\ f\ _{B_q^\alpha(L_q)} = 125.14$, correlation = -0.96						
179.0550	210.3713	88.4807	130.2817	70.0330	114.4309	no
170.9440	280.5647	82.5662	182.3281	64.5305	162.1056	no
162.4270	332.4486	76.1940	221.7319	58.8987	200.5562	yes
153.4390	347.2902	69.2377	233.1997	53.2422	217.1064	yes
143.8900	329.2055	61.4996	228.5820	47.7693	226.2830	yes
133.6610	342.1875	52.6359	229.8249	42.8144	241.9251	no
img0002: $\alpha = 0.4540$, $\ f\ _{B_q^\alpha(L_q)} = 33.10$, correlation = -0.98						
179.0550	53.2541	99.5781	39.4571	83.3562	36.7997	yes
170.9440	65.2335	95.0225	50.6094	78.9509	47.9768	yes
162.4270	75.2916	90.2371	58.7261	74.4123	56.5222	yes
153.4390	84.0526	85.1833	64.6977	69.7469	63.1737	yes
143.8900	101.4971	79.8102	75.8258	64.9754	74.9687	yes
133.6610	156.7822	74.0482	96.3182	60.1433	95.0280	yes
img0003: $\alpha = 0.5901$, $\ f\ _{B_q^\alpha(L_q)} = 63.55$, correlation = -1.00						
179.0550	57.4298	100.6510	38.2988	81.0953	34.7466	yes
170.9440	85.2939	95.2735	56.6259	75.7603	51.2990	yes
162.4270	120.6397	89.5736	80.7629	70.2193	73.2333	yes
153.4390	164.5886	83.4855	112.0940	64.4815	102.0418	yes
143.8900	208.4330	76.9170	140.1552	58.5888	131.0985	yes
133.6610	270.8577	69.7324	184.1014	52.6496	175.5448	yes
img0004: $\alpha = 0.5528$, $\ f\ _{B_q^\alpha(L_q)} = 60.88$, correlation = -1.00						
179.0550	64.7455	98.5727	44.2242	79.8460	40.5783	yes
170.9440	92.4745	93.3053	62.5352	74.6833	57.4896	yes
162.4270	130.4371	87.7222	86.7435	69.3357	79.8820	yes
153.4390	176.4368	81.7588	118.8742	63.8189	108.5508	yes
143.8900	227.2213	75.3247	157.5147	58.1825	146.8503	yes
133.6610	275.7021	68.2870	188.0698	52.5418	179.1268	yes
img0005: $\alpha = 0.7437$, $\ f\ _{B_q^\alpha(L_q)} = 296.55$, correlation = -0.97						
179.0550	228.8171	91.8893	133.0216	69.5146	109.1299	no
170.9440	345.7875	85.0457	207.5407	62.8672	170.2372	no
162.4270	472.8143	77.6008	290.4874	55.9800	239.8775	no
153.4390	571.6021	69.3614	350.2825	49.0033	293.5128	no
143.8900	627.1891	60.0011	372.4124	42.3075	328.3381	yes
133.6610	600.4391	48.8800	368.2965	36.5452	358.3319	yes
img0006: $\alpha = 0.5390$, $\ f\ _{B_q^\alpha(L_q)} = 91.93$, correlation = -0.96						
179.0550	150.0127	91.7629	96.5456	73.5256	85.6978	no
170.9440	196.4244	86.1745	134.0817	68.2293	121.3433	yes
162.4270	229.5340	80.1976	157.2531	62.7915	144.9852	yes
153.4390	247.4985	73.7379	169.2620	57.2359	159.4247	yes
143.8900	260.3081	66.6550	172.4382	51.7283	171.0568	no
133.6610	304.1729	58.7240	185.3874	46.4951	192.6011	no
img0007: $\alpha = 0.6814$, $\ f\ _{B_q^\alpha(L_q)} = 123.45$, correlation = -0.99						
179.0550	95.0303	98.5091	56.1827	77.1839	48.4237	yes
170.9440	151.2305	92.4849	92.7667	71.1903	79.6520	yes
162.4270	221.1551	86.0399	142.9906	64.9442	123.9949	yes
153.4390	281.5109	79.0713	191.8522	58.4764	170.6784	yes
143.8900	328.3248	71.4260	227.5274	51.8928	207.7471	yes
133.6610	346.1154	62.8576	253.6694	45.4539	244.6782	yes
img0008: $\alpha = 0.6030$, $\ f\ _{B_q^\alpha(L_q)} = 250.50$, correlation = -0.98						
179.0550	341.9449	82.4756	186.9621	63.2527	159.9513	no
170.9440	486.4462	75.7248	260.5653	57.2482	221.6034	no
162.4270	631.5497	68.3101	329.4598	51.2253	282.1386	no
153.4390	739.8782	59.9857	377.8892	45.4529	335.0769	no
143.8900	781.4032	50.3021	398.6313	40.3575	376.4353	yes
133.6610	769.5566	38.2403	412.2842	36.3601	412.1506	yes

We then added i.i.d. Gaussian noise with standard deviation 32 to the pixels of each image f to obtain noisy images \tilde{f} . Wavelet shrinkage with various parameters λ was applied to remove noise from each image and obtain noise-removed images \tilde{f}_λ , and the error was measured.

Table 1 contains the results of our tests, which are reported using slightly different conventions than elsewhere in this paper. The listing for each image contains six lines, corresponding to the six sizes of our downsampled images, with the largest (original) image listed first. We calculated three values of the shrinkage parameter. The first, which does not depend on image smoothness parameters, is the one proposed by Donoho and Johnstone in VisuShrink,

$$\lambda_V = 32\sqrt{2\log M},$$

where the image has M pixels ($M = 6,291,456$ for the largest image, etc.). Beginning with the largest image, we have $\lambda_V = 179.0550, 170.9440, 162.4270, 153.4390, 143.8900$, and $\lambda_V = 133.6610$, respectively. The actual shrinkage parameter is λ_V/\sqrt{M} , but we thought comparisons would be easier if we left out the factor \sqrt{M} in the table. In all cases, the shrinkage parameters are multiples of the standard deviation of the noise, which is 32.

We also report

$$E_V := \|f - \tilde{f}_{\lambda_V}\|_{L_2(I)}^2,$$

the square of the $L_2(I)$ error after wavelet shrinkage. We square the error before reporting it to make comparison with our bound on $E(\|f - \tilde{f}_\lambda\|_{L_2(I)}^2)$ easier. The error is normalized so that the error before shrinkage, $\|f - \tilde{f}\|_{L_2(I)}^2$, is almost exactly 1024.

In addition to λ_V , we calculated

$$\lambda_e = 32a,$$

where a was calculated from (32). This “easy” estimate does not minimize (30), but does take into account the smoothness parameter α and the signal-to-noise ratio $\|f\|_{B_q^\alpha(L_q(I))}/\sigma$. E_e is the error after shrinkage with λ_e/\sqrt{M} .

Finally, we found the a that minimized the bound (30), and calculated the “critical” parameter

$$\lambda_c = 32a.$$

Since this calculation is trivial in practice, there is really no reason to use λ_e except for illustrative purposes. We report the error for this critical λ as E_c .

We resist calling λ_c the “optimal” λ . There is an optimal parameter λ_o that does minimize the error $\|f - \tilde{f}_{\lambda_o}\|_{L_2(I)}^2$. We estimated λ_o as follows. We calculated the error after shrinkage by the parameters $\lambda_c, .9\lambda_c$, and $1.1\lambda_c$, and fitted a quadratic polynomial $E(\lambda)$ through these three points. We found the value of λ at which $E(\lambda)$ was a minimum, and called this value λ_o . For each image in the table, we reported whether λ_o was between $.9\lambda_c$ and $1.1\lambda_c$, i.e., whether the optimal λ was within 10% of the critical λ

TABLE 1 (CONTINUED)
Shrinkage Parameters and Errors

λ_V	E_V	λ_e	E_e	λ_c	E_c	$\frac{ \lambda_c - \lambda_o }{\lambda_c} \leq .1$
img0009: $\alpha = 0.4379$, $\ f\ _{B_q^\alpha(L_q)} = 48.91$, correlation = -1.00						
179.0550	85.2700	92.4943	54.3108	77.0353	51.4088	yes
170.9440	119.5252	87.6959	73.8423	72.5615	68.6422	yes
162.4270	162.5655	82.6192	103.3199	67.9773	96.0030	yes
153.4390	206.4507	77.2094	137.5044	63.3117	128.8130	yes
143.8900	246.6295	71.3908	163.2856	58.6226	155.9003	yes
133.6610	297.6367	65.0539	198.9910	54.0130	194.8449	yes
img0010: $\alpha = 0.4516$, $\ f\ _{B_q^\alpha(L_q)} = 49.89$, correlation = -1.00						
179.0550	86.8665	93.3829	55.0300	77.4584	51.3298	yes
170.9440	120.0591	88.5277	76.8146	72.8945	70.9642	yes
162.4270	157.7757	83.3903	104.6890	68.2101	97.2142	yes
153.4390	199.4703	77.9148	136.7445	63.4328	128.4150	yes
143.8900	230.5630	72.0243	157.7135	58.6193	151.1259	yes
133.6610	324.0132	65.6070	200.2188	53.8740	191.8540	yes
img0011: $\alpha = 0.5508$, $\ f\ _{B_q^\alpha(L_q)} = 86.41$, correlation = -0.98						
179.0550	126.5675	93.6162	84.6573	75.0308	76.0954	no
170.9440	170.6121	88.0656	115.7660	69.7075	104.5577	yes
162.4270	214.3538	82.1409	145.2083	64.2144	132.1884	yes
153.4390	252.9923	75.7543	168.0925	58.5979	154.6621	yes
143.8900	292.9337	68.7770	196.1860	52.9671	183.8881	yes
133.6610	340.4842	61.0070	229.5527	47.5358	222.6496	yes
img0012: $\alpha = 0.5474$, $\ f\ _{B_q^\alpha(L_q)} = 61.84$, correlation = -1.00						
179.0550	63.6581	97.9672	42.4497	79.3720	39.5742	yes
170.9440	96.5165	92.6996	61.6802	74.2251	57.3486	yes
162.4270	138.8406	87.1142	86.5147	68.8967	79.1298	yes
153.4390	184.5532	81.1451	123.8349	63.4050	113.8862	yes
143.8900	226.5512	74.7007	152.6660	57.8037	144.7776	yes
133.6610	298.3634	67.6450	187.3000	52.2137	182.0770	yes
img0013: $\alpha = 0.5255$, $\ f\ _{B_q^\alpha(L_q)} = 141.98$, correlation = -0.94						
179.0550	317.4228	83.9277	191.9675	66.5094	167.0649	no
170.9440	402.3190	77.8834	254.3060	61.1206	224.0564	no
162.4270	443.8249	71.3287	285.0475	55.6787	255.3935	no
153.4390	437.1571	64.1074	282.5531	50.3389	262.9107	yes
143.8900	407.0308	55.9618	268.2220	45.3643	264.1632	yes
133.6610	404.4600	46.4079	262.5576	41.0896	269.7422	no
img0014: $\alpha = 0.5956$, $\ f\ _{B_q^\alpha(L_q)} = 115.23$, correlation = -0.98						
179.0550	147.4220	93.1613	93.6570	73.5703	82.1444	no
170.9440	209.0038	87.2882	135.8491	67.9142	119.3244	no
162.4270	269.9440	80.9902	180.4331	62.0694	159.2547	no
153.4390	314.1836	74.1593	216.1746	56.0993	195.2496	yes
143.8900	335.4080	66.6318	228.4760	50.1557	216.6789	yes
133.6610	394.1094	58.1376	242.9567	44.5354	242.1689	no
img0015: $\alpha = 0.6630$, $\ f\ _{B_q^\alpha(L_q)} = 118.93$, correlation = -1.00						
179.0550	88.4063	97.7155	59.8222	76.7231	53.7318	yes
170.9440	134.2533	91.7411	85.7379	70.8069	75.8413	yes
162.4270	197.5251	85.3495	120.6931	64.6493	104.5225	yes
153.4390	302.5844	78.4389	175.1750	58.2848	149.3235	yes
143.8900	412.6219	70.8574	241.8167	51.8237	202.7265	no
133.6610	470.7020	62.3609	281.4373	45.5270	251.7184	yes
img0016: $\alpha = 0.4545$, $\ f\ _{B_q^\alpha(L_q)} = 43.88$, correlation = -0.97						
179.0550	83.4436	95.5487	60.7350	79.4413	56.4910	yes
170.9440	100.8842	90.7874	76.2909	74.9135	72.0715	yes
162.4270	111.3879	85.7621	86.0488	70.2575	82.8485	yes
153.4390	120.8698	80.4235	91.2814	65.4918	89.6117	yes
143.8900	138.0214	74.7043	100.5314	60.6586	100.9561	yes
133.6610	199.8825	68.5093	131.4323	55.8399	131.1338	yes

TABLE 1 (CONTINUED)
Shrinkage Parameters and Errors

λ_V	E_V	λ_e	E_e	λ_c	E_c	$\frac{ \lambda_c - \lambda_o }{\lambda_c} \leq .1$
img0017: $\alpha = 0.4955$, $\ f\ _{B_q^\alpha(L_q)} = 65.06$, correlation = -1.00						
179.0550	114.2074	93.1647	76.9241	75.9970	70.4480	no
170.9440	158.4131	87.9713	102.3399	71.0754	92.3259	no
162.4270	213.8878	82.4514	140.5172	66.0129	125.2978	no
153.4390	259.3381	76.5344	181.7903	60.8467	163.9227	no
143.8900	292.2991	70.1199	212.8323	55.6572	198.8557	yes
133.6610	309.9920	63.0562	220.8216	50.5964	211.9355	yes
img0018: $\alpha = 0.4725$, $\ f\ _{B_q^\alpha(L_q)} = 79.97$, correlation = -0.98						
179.0550	183.0025	87.9730	117.8489	71.7434	105.2940	no
170.9440	231.7082	82.6334	153.1518	66.8706	137.2509	no
162.4270	267.6379	76.9240	184.2612	61.9020	167.4996	no
153.4390	280.9737	70.7555	193.6981	56.9075	179.3467	yes
143.8900	283.8122	63.9951	198.3832	52.0163	190.4115	yes
133.6610	318.9578	56.4305	203.0324	47.4341	199.6398	yes
img0019: $\alpha = 0.5622$, $\ f\ _{B_q^\alpha(L_q)} = 97.43$, correlation = -0.99						
179.0550	124.9804	92.8377	77.9571	74.0073	70.0833	yes
170.9440	177.7580	87.1613	112.9656	68.5756	101.2258	yes
162.4270	229.4604	81.0884	147.5607	62.9730	133.4492	yes
153.4390	265.8489	74.5223	171.0784	57.2558	158.0337	yes
143.8900	286.2085	67.3189	183.9823	51.5544	176.3429	yes
133.6610	333.3324	59.2459	210.2506	46.1183	213.3032	no
img0020: $\alpha = 0.4727$, $\ f\ _{B_q^\alpha(L_q)} = 59.85$, correlation = -1.00						
179.0550	98.1254	92.4599	65.0133	75.9604	60.1195	yes
170.9440	138.6606	87.3934	85.8788	71.1997	77.7485	yes
162.4270	190.5754	82.0145	117.6737	66.3124	105.4753	no
153.4390	246.1533	76.2571	151.7334	61.3348	136.5410	no
143.8900	326.1522	70.0279	187.7587	56.3424	171.0915	yes
133.6610	437.1484	63.1876	258.9076	51.4726	237.8797	yes
img0021: $\alpha = 0.5278$, $\ f\ _{B_q^\alpha(L_q)} = 91.12$, correlation = -0.99						
179.0550	143.3057	90.9459	91.0350	73.0231	81.9721	yes
170.9440	193.9839	85.3831	125.3406	67.7863	113.2546	yes
162.4270	239.1111	79.4317	158.1172	62.4069	144.6825	yes
153.4390	284.4246	72.9967	183.8413	56.9484	170.9458	yes
143.8900	328.7246	65.9367	207.3957	51.5450	198.7390	yes
133.6610	405.1234	58.0239	245.8031	46.4348	243.2554	yes
img0022: $\alpha = 0.4996$, $\ f\ _{B_q^\alpha(L_q)} = 61.72$, correlation = -0.98						
179.0550	98.2154	94.2705	66.6163	76.9527	61.1441	yes
170.9440	131.0469	89.1128	90.4218	72.0333	83.7140	yes
162.4270	161.2662	83.6376	113.0111	66.9666	105.4138	yes
153.4390	180.6328	77.7779	128.7200	61.7837	122.9030	yes
143.8900	211.5354	71.4392	147.9017	56.5554	144.5637	yes
133.6610	266.3566	64.4804	177.4961	51.4196	178.0597	yes
img0023: $\alpha = 0.6367$, $\ f\ _{B_q^\alpha(L_q)} = 67.91$, correlation = -1.00						
179.0550	46.6726	102.9010	29.5961	82.4312	26.5648	yes
170.9440	74.6494	97.3866	47.6397	76.8953	42.2607	yes
162.4270	112.5904	91.5409	73.6272	71.1298	65.4091	yes
153.4390	163.7535	85.2955	105.5340	65.1372	93.9466	yes
143.8900	229.7280	78.5551	145.3447	58.9506	129.1398	yes
133.6610	316.4717	71.1792	202.5109	52.6708	187.4058	yes
img0024: $\alpha = 0.6804$, $\ f\ _{B_q^\alpha(L_q)} = 195.87$, correlation = -0.96						
179.0550	200.2875	92.5502	120.7357	71.2732	101.5761	no
170.9440	282.2606	86.1160	179.3111	65.0360	152.5866	no
162.4270	350.0949	79.1605	228.3692	58.5769	197.4706	no
153.4390	392.0606	71.5319	250.1404	52.0003	222.1221	yes
143.8900	432.8792	62.9859	258.2607	45.5631	243.1299	yes
133.6610	521.6043	53.0815	293.8169	39.7510	285.7367	no



FIG. 1. *Original image.*



FIG. 3. *Noise removed with $\lambda = \sigma_0 2^{-m} \sqrt{2 \ln 2^{2m}}$ (VisuShrink).*



FIG. 2. *Original image with noise, $\sigma_0 = 32$.*



FIG. 4. *Noise removed by minimizing (30).*

calculated by minimizing (30). The results show that the optimal λ was within 10% of λ_c in 106 of 144 cases.

As a final example, we compare VisuShrink, our method, and SureShrink on an image with extreme smoothness properties—a 512 by 512 section of fp1.pgm, the first test image of the FBI wavelet compression algorithm, available at <ftp://ftp.c3.lanl.gov/pub/WSQ>. The original image is displayed in Figure 1. We note first that although the image is rather smooth (there is a great deal of structure, but no texture and no real

edges), it contains a *lot* of information. Thus, we expect $f \in B_q^\alpha(L_q(I))$, $1/q = \alpha/2 + 1/2$, with rather high α , but also high $\|f\|_{B_q^\alpha(L_q(I))}$. We compressed this image at several compression levels, with a quantization strategy that attempts to minimize the error in $L_2(I)$, and obtained for $N = 162159, 111957, 66057, 33952, 17215,$ and 8262 coefficients $L_2(I)$ errors of 1.1873394, 2.1595381, 3.7883904, 6.2393051, 9.6140564, and 14.3311631 grey scales, respec-



FIG. 5. *Noise removed using SureShrink.*



FIG. 6. *Noise removed with best linear method.*

tively. Thus, we have

$$\|f - f_N\|_{L_2(I)} \approx 24504.6 N^{-1.61466/2}.$$

Thus, $f \in B_q^\alpha(L_q(I))$ with $\alpha = 1.61466$ and $q = .76492$, i.e., f has relatively high order of smoothness, but $\|f\|_{B_q^\alpha(L_q(I))} \approx 24504.6$, i.e., in this space the norm is very high. The correlation coefficient of this line (on a log-log graph) is -0.982898 .

In Figure 2, we show the same image with i.i.d. Gaussian noise with mean zero and standard deviation 32 added

to each pixel. Let us denote the original pixels by p_j , and their average over the entire image by \bar{P} . Then, a signal to noise ratio for this image can be defined by

$$\sqrt{\frac{\sum_j (p_j - \bar{P})^2}{262,144}} \div 32 = \frac{49.6414327}{32} = 1.55.$$

In other words, the standard deviation of the signal is only about 1.55 times the standard deviation of the noise we added—this signal-to-noise ratio is quite small, and this number leads us to expect the signal to be almost obliterated by the added noise.

Obviously, this is nonsense. The added noise obscures hardly any information in the image at all—under any reasonable measure, the signal-to-noise ratio of the noisy image is extremely high. Our new definition (33) of signal-to-noise ratio gives

$$\frac{\|F\|_{B_q^\alpha(L_q(I))}}{\sigma_0} = \frac{24504.6}{32} = 765.8,$$

which predicts that the noise hardly affects the visual perception of the signal at all.

Figure 3 shows the result of setting $\lambda = \lambda_V$, the VisuShrink value of $\sqrt{2 \ln 262,144} = 159.850$. Many image features are smoothed away. The root-mean-square error between the original and smoothed images is 26.8321869 grey scales. We did not use the VisuShrink procedure in Wavelab, developed by Donoho and others, because their program is designed to handle images only of dimension $2^m \times 2^m$ for some m , and our images are rectangular, not square. We did compare their program to ours for the next example, and obtained essentially the same results, both visually and in the measured error—only the location of the artifacts was different.

Figure 4 shows the result of minimizing (30), which gives $\lambda_c = 43.516416$ and leads to an RMS error between the original and smoothed images of 17.0109976 grey scales. The RMS error when shrinking by $0.9\lambda_c$ is 17.0305316, while the RMS error when shrinking by $1.1\lambda_c$ is 17.1634679, so we estimate that the optimal λ for this problem is indeed within 10% of λ_c . Plugging this value of a into (30) gives an upper bound for the expected RMS error of 18.4938542 grey scales, which is fairly close to the real error.

Figure 5 shows the results of applying Michael Hilton's implementation of SureShrink, available at http://www.cs.sc Carolina.edu/ABOUT_US/Faculty/Hilton/shrink-demo.html. SureShrink uses a different shrinkage parameter for each dyadic level (indeed, for each type of wavelet at each dyadic level). SureShrink leaves more noise in the smoothed image, but removes fewer image details. It achieves an RMS error of 13.3632283 grey scales, significantly better than VisuShrink or our method with the critical (or even optimal) λ . We leave it to the reader to compare the visual quality of the two de-noised images.

Finally, Figure 6 presents the results of applying the best linear method of Section 8, which removes all wavelet

terms in the two highest dyadic levels. The RMS error is 16.9924173, essentially the same error as the nonlinear method with the critical $\lambda = \lambda_c$. The nonlinear method achieves significantly better visual results, however.

We believe that these results show several things. Minimizing our bound on the error (30) leads to near-optimal shrinkage parameters for noise removal with wavelet shrinkage when using a single shrinkage parameter for all dyadic levels. Our technique for estimating the smoothness of images leads to accurate estimates of the true smoothness of images. The performance of wavelet image processing algorithms can be predicted accurately using only the two smoothness parameters α and $\|f\|_{B_q^\alpha(L_q(I))}$. Our new definition (33) of a signal-to-noise ratio is a better measure than the one typically used. SureShrink achieved better results on our rather extreme example than any wavelet shrinkage method that uses a single shrinkage parameter over all dyadic levels. Finally, even though there is no *a priori* reason to assume that the smoothness of real images can be characterized by only these two parameters, and even though it is easy to come up with images that do *not* satisfy this assumption (a montage of unrelated images, for example), in practice real images often have rather uniform smoothness over the entire image.

Appendix I

In this Appendix we prove the assertion at the end of Section 3. Since we use arguments from [22], we adopt their notation in this section. Thus, we assume that functions f and g in $L_2(I)$ have orthogonal wavelet expansions

$$f = \sum_{j,k,\psi} c_{j,k,\psi} \psi_{j,k} \quad \text{and} \quad g = \sum_{j,k,\psi} d_{j,k,\psi} \psi_{j,k}$$

and consider the minimization problem

$$(38) \quad \frac{1}{2\lambda} \|f - g\|_{L_2(I)}^2 + J(g)$$

where

$$J(g) := \|g\|_{B_\infty^1(L_1(I))} := \sup_k \sum_{j,\psi} |d_{j,k,\psi}|.$$

We define $\hat{g} = \sum_{j,k,\psi} \hat{d}_{j,k,\psi} \psi_{j,k}$ with

$$\hat{d}_{j,k,\psi} = \begin{cases} 0, & |c_{j,k,\psi}| \leq \lambda_k, \\ c_{j,k,\psi} - \lambda_k \operatorname{sgn}(c_{j,k,\psi}), & |c_{j,k,\psi}| > \lambda_k, \end{cases}$$

to be the solution of

$$\min_g \|f - g\|_{L_2(I)}, \quad \text{subject to } J(g) \leq M.$$

We let

$$\lambda := \sum_k \lambda_k,$$

which is finite since

$$(39) \quad \begin{aligned} \infty > \langle \hat{g}, f - \hat{g} \rangle &= \sum_{j,k,\psi} \hat{d}_{j,k,\psi} (c_{j,k,\psi} - \hat{d}_{j,k,\psi}) \\ &= \sum_{j,k,\psi} \lambda_k |\hat{d}_{j,k,\psi}| \\ &= \sum_k \lambda_k \sum_{j,\psi} |\hat{d}_{j,k,\psi}| = M \sum_k \lambda_k. \end{aligned}$$

The second equality holds because $\hat{d}_{j,k,\psi}$ is nonzero only when $|c_{j,k,\psi}| > \lambda_k$, in which case $\hat{d}_{j,k,\psi} = c_{j,k,\psi} - \operatorname{sgn}(c_{j,k,\psi})\lambda_k$ and $|\hat{d}_{j,k,\psi}| = |c_{j,k,\psi}| - \lambda_k$. The last equality holds since $\sum_{j,\psi} |\hat{d}_{j,k,\psi}| = M$ whenever λ_k is nonzero. We show that if \hat{g} is the minimizer over g of (38) where $\lambda = \sum_k \lambda_k$. This is equivalent to the claim made at the end of Section 3. Notice that if $J(f) \leq M$, then $\hat{g} = f$ and $\lambda = 0$, so that $\frac{1}{2\lambda} = +\infty$ and the claim is clearly true. We will thus assume that $J(f) > M$, in this case, the λ_k are chosen in a way that ensures that $J(\hat{g}) = M$.

We consider the closed, convex set V defined by

$$(40) \quad J^*(v) := \sup_{u \in L_2(I)} [\langle v, u \rangle - J(u)] = \begin{cases} 0, & x \in V, \\ \infty, & x \notin V, \end{cases}$$

In convex analysis (see [22]), J^* is called the Legendre-Fenchel conjugate of J , and the function on the right of (40) is called the characteristic function of V and denoted by χ_V . (This conflicts with the usual definition

$$\chi_V(x) = \begin{cases} 1, & x \in V, \\ 0, & x \notin V, \end{cases}$$

which is not used in this Appendix.) It is a standard result that the Legendre-Fenchel conjugate of a convex, homogeneous function is the characteristic function of a closed convex set.

We now characterize the set V . For any g we have that

$$J(g) = \sup_k \sum_{j,\psi} |d_{j,k,\psi}| = \sup_k \sum_{j,\psi} |\langle g, \psi_{j,k} \rangle|;$$

thus, since $\ell_1^* = \ell_\infty$, $J(g) = \sup_{w \in W} \langle g, w \rangle$, where

$$W = \{w = \sum_{j,k,\psi} t_{j,k,\psi} \psi_{j,k} \mid \sum_k \sup_{j,\psi} |t_{j,k,\psi}| \leq 1\} \cap L_2(I).$$

In other words, $J = \chi_W^*$, and $J^* = \chi_V = (\chi_W)^{**}$ is the convex lower-semi-continuous envelope of the characteristic function of W . This implies that V is the closed, convex envelope of W in $L_2(I)$, or, since W is itself closed in $L_2(I)$ and convex, $V = W$.

It is now clear that $\frac{f - \hat{g}}{\lambda}$ belongs to V , so that

$$J^*\left(\frac{f - \hat{g}}{\lambda}\right) = 0,$$

thus using the fact that $J(\hat{g}) = M$ and (39), we deduce

$$J(\hat{g}) + J^*\left(\frac{f - \hat{g}}{\lambda}\right) = M = \left\langle \frac{f - \hat{g}}{\lambda}, \hat{g} \right\rangle.$$

Since for any convex lower-semi-continuous function J ,

$$\begin{aligned} v \in \partial J(u) &\iff u \in \partial J^*(v) \\ &\iff \langle v, u \rangle = J^*(v) + J(u), \end{aligned}$$

where ∂J (resp., ∂J^*) denotes the subdifferential of J (resp., J^*) (see [22]), we deduce

$$\frac{f - \hat{g}}{\lambda} \in \partial J(\hat{g}) \iff \partial J(\hat{g}) + \frac{\hat{g} - f}{\lambda} \ni 0$$

that is to say, \hat{g} , is the (unique) minimizer of (38).

Notice that by Legendre-Fenchel duality we can show that as minimizer of (38), \hat{g} satisfies

$$\hat{g} = f - \Pi_{\lambda V} f,$$

where $\Pi_{\lambda V}$ denotes the $L_2(I)$ projection onto the set λV .

Appendix II

In this Appendix we prove the assertions made at the end of Section 4. We use the notation of that section.

The first claim (19) is that

$$\|f_m - f_o\|_{L_2(I)} \leq C2^{-\alpha m} \|F\|_{W^\alpha(L_2(I))}.$$

Note that

$$\|f_m - f_o\|_{L_2(I)}^2 \leq C \sum_j |\langle F, \tilde{\phi}_{j,m} \rangle - \langle F, \tilde{\Phi}_{j,m} \rangle|^2$$

since $\{\phi_{j,m}\}_j$ is a Riesz basis for $\text{span}\{\phi_{j,m}\}_j$. Let $P_{j,m}$ be the polynomial of (total) degree $< s$ of best L_2 approximation to F on $I_{j,m}$. Since $\langle P_{j,m}, \tilde{\phi}_{j,m} \rangle = \langle P_{j,m}, \tilde{\Phi}_{j,m} \rangle$ by (18),

$$\begin{aligned} \|f_m - f_o\|_{L_2(I)}^2 &\leq C \sum_j |\langle F, \tilde{\phi}_{j,m} \rangle - \langle F, \tilde{\Phi}_{j,m} \rangle|^2 \\ &= C \sum_j |\langle F - P_{j,m}, \tilde{\phi}_{j,m} \rangle - \langle F - P_{j,m}, \tilde{\Phi}_{j,m} \rangle|^2 \\ &\leq C \sum_j [|\langle F - P_{j,m}, \tilde{\phi}_{j,m} \rangle|^2 + |\langle F - P_{j,m}, \tilde{\Phi}_{j,m} \rangle|^2] \\ &\leq C \sum_j \int_{\tilde{I}_{j,m}} |F - P_{j,m}|^2 \end{aligned}$$

where $\tilde{I}_{j,m}$ is the smallest interval containing the supports of both $\tilde{\phi}_{j,m}$ and $\tilde{\Phi}_{j,m}$, since, by our assumptions, $\max_j \|\tilde{\phi}_{j,m}\|_{L_2(I)}$ and $\max_j \|\tilde{\Phi}_{j,m}\|_{L_2(I)}$ are both bounded by a constant. Now, because the diameter of $\tilde{I}_{j,m}$ is bounded by $C2^{-m}$,

$$\int_{\tilde{I}_{j,m}} |F - P_{j,m}|^2 \leq C \mathbf{w}_s(F, C2^{-2m})_{L_2(\tilde{I}_{j,m})}^2,$$

where $\mathbf{w}_s(f, t)_{L_p(J)}$ is the averaged modulus of smoothness; see [9]. Since the averaged modulus of smoothness is subadditive on sets, and each $x \in I$ is contained in at most

C_0 intervals $\tilde{I}_{j,m}$, with C_0 an absolute constant, we have

$$\begin{aligned} \|f_m - f_o\|_{L_2(I)}^2 &\leq C \sum_j \mathbf{w}_s(F, C2^{-2m})_{L_2(\tilde{I}_{j,m})}^2 \\ &\leq C \mathbf{w}_s(F, C2^{-2m})_{L_2(I)}^2 \\ &\leq C \omega_s(F, 2^{-2m})_{L_2(I)}^2 \end{aligned}$$

where $\omega_s(F, 2^{-2m})_{L_2(I)}$ is the usual modulus of smoothness. The claim now follows from the fact that

$$\omega_s(F, 2^{-2m})_{L_2(I)} \leq C2^{-\alpha m} \|F\|_{W^\alpha(L_2(I))}$$

when $\alpha < s$.

To prove the second assertion (20), we expand

$$\begin{aligned} f_o &= \sum_j \langle F, \tilde{\Phi}_{j,m} \rangle \phi_{j,m} \\ &= \sum_j \langle f_o, \tilde{\phi}_{j,0} \rangle \phi_{j,0} + \sum_{0 \leq k < m} \sum_j \langle f_o, \tilde{\psi}_{j,k} \rangle \psi_{j,k} \end{aligned}$$

and the question is whether

$$\begin{aligned} \|f_o\|_{B_p^\alpha(L_p(I))}^p &\asymp \left\| \sum_j \langle f_o, \tilde{\phi}_{j,0} \rangle \phi_{j,0} \right\|_{L_p(I)}^p \\ \text{(A)} \quad &+ \sum_{0 \leq k < m} \sum_{j,\psi} 2^{\alpha k p} \|\langle f_o, \tilde{\psi}_{j,k} \rangle \psi_{j,k}\|_{L_p(I)}^p \end{aligned}$$

is bounded by a constant times $\|F\|_{B_p^\alpha(L_p(I))}^p$. We shall bound the two sums on the right side of (A) separately.

We begin with the second. Each coefficient of $\psi_{j,k}$ can be written

$$\begin{aligned} \langle f_o, \tilde{\psi}_{j,k} \rangle &= \left\langle \sum_\ell \langle F, \tilde{\Phi}_{\ell,m} \rangle \phi_{\ell,m}, \tilde{\psi}_{j,k} \right\rangle \\ &= \left\langle F, \sum_\ell \langle \phi_{\ell,m}, \tilde{\psi}_{j,k} \rangle \tilde{\Phi}_{\ell,m} \right\rangle \\ &= \langle F, \tilde{\psi}_J \rangle, \end{aligned}$$

where

$$\tilde{\psi}_J = \sum_\ell \langle \phi_{\ell,m}, \tilde{\psi}_{j,k} \rangle \tilde{\Phi}_{\ell,m},$$

is associated with the interval $J = [j/2^k, (j+1)/2^k]$ (in one dimension).

We use Theorem 3.7 of Frazier and Jawerth's paper on the ϕ transform [23] to show that

$$\sum_{0 \leq k < m} \sum_{j,\psi} 2^{\alpha k p} \|\langle f_o, \tilde{\psi}_{j,k} \rangle \psi_{j,k}\|_{L_p(I)}^p \leq C \|F\|_{B_p^\alpha(L_p(I))}^p.$$

This will follow if we show that ψ_J satisfies Conditions (3.7) to (3.10) of that work. Condition (3.8), a size and decay condition on ψ_J , is satisfied because $\tilde{\Phi}_{j,m}$ satisfies the first and second assumptions on page 8, and the diameter of the support of $\tilde{\psi}_J$ is bounded by $C2^{-k}$ whenever $k < m$. Conditions (3.9) and (3.10) are vacuous (automatically satisfied) when $1/p < \alpha/d + 1$ in d dimensions. It remains to show (3.7), which is

$$\langle x^s, \psi_J(x) \rangle = 0 \quad \text{for } |s| \leq [\alpha],$$

where $[\alpha]$ is the integer part of α , $\alpha - 1 < [\alpha] \leq \alpha$.

But this condition is trivially satisfied by our construction, because

$$\langle P, \tilde{\psi}_J \rangle = \left\langle \sum_{\ell} \langle P, \tilde{\Phi}_{\ell,m} \rangle \phi_{\ell,m}, \tilde{\psi}_{j,k} \right\rangle = 0$$

since

$$\sum_{\ell} \langle P, \tilde{\Phi}_{\ell,m} \rangle \phi_{\ell,m}$$

is a polynomial of degree $\leq \lfloor \alpha \rfloor < r$ whenever P is.

Finally, we proceed to bound the first term on the right-hand side of (A). Recall that $B_p^\alpha(L_p(I))$ is embedded in $L_q(I)$, for some $q > 1$ whenever $1/p < \alpha/d + 1$ in d dimensions. It is not too hard to see that the operator \mathbf{P} that takes F to

$$\sum_j \langle f_o, \tilde{\phi}_{j,0} \rangle \phi_{j,0} = \sum_j \left\langle \sum_{\ell} \langle F, \tilde{\Phi}_{\ell,m} \rangle \phi_{\ell,m}, \tilde{\phi}_{j,0} \right\rangle \phi_{j,0}$$

is a bounded operator from $L_q(I)$ to the space $V_0 := \text{span}\{\phi_{j,0}\}$ with the same norm. Hence,

$$\begin{aligned} \left\| \sum_j \langle f_o, \tilde{\phi}_{j,0} \rangle \phi_{j,0} \right\|_{L_p(I)} &\leq C \left\| \sum_j \langle f_o, \tilde{\phi}_{j,0} \rangle \phi_{j,0} \right\|_{L_q(I)} \\ &\leq C \|F\|_{L_q(I)} \\ &\leq C \|F\|_{B_p^\alpha(L_p(I))}. \end{aligned}$$

Here the first inequality follows because V_0 is finite-dimensional, so all (quasi-)norms are equivalent; the second inequality follows because \mathbf{P} is bounded on $L_q(I)$; and the third inequality follows from the embedding of $B_p^\alpha(L_p(I))$ into $L_q(I)$.

REFERENCES

- [1] C. Bouman and K. Sauer, *Bayesian estimation of transmission tomograms using segmentation based optimization*, IEEE Trans. Nuclear Science, 39, 4 (1992), pp. 1144–1152.
- [2] G. Chavent and K. Kunisch, *Convergence of Tikhonov regularization for constrained ill-posed problems*, Inverse Problems, 10 (1994), pp. 63–76.
- [3] A. Cohen, I. Daubechies, and J.-C. Feauveau, *Biorthogonal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 45 (1992), pp. 485–560.
- [4] A. Cohen, I. Daubechies, and P. Vial, *Wavelets on the interval and fast wavelet transforms.*, Appl. Comput. Harmonic Analysis, 1 (1993), pp. 54–81.
- [5] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics 91, SIAM, Philadelphia, 1992.
- [6] R. DeVore, B. Jawerth, and B. Lucier, *Image compression through wavelet transform coding*, IEEE Trans. Information Theory, 38, 2 (1992), pp. 719–746, Special issue on Wavelet Transforms and Multiresolution Analysis.
- [7] ———, *Surface compression*, Computer Aided Geom. Design, 9 (1992), pp. 219–239.
- [8] R. DeVore, B. Jawerth, and V. Popov, *Compression of wavelet decompositions*, Amer. J. Math., 114 (1992), pp. 737–785.
- [9] R. DeVore and G. Lorentz, *Constructive Approximation*, Springer-Verlag, New York, 1993.
- [10] R. A. DeVore and B. J. Lucier, *Classifying the smoothness of images: Theory and applications to wavelet image processing*, in ICIP-94: Proceedings of the 1994 IEEE International Conference on Image Processing, Austin, TX, November 13–16 II, IEEE Press, Los Alamitos, CA, 1994, pp. 6–10.
- [11] ———, *Error bounds for image compression by zero-tree coding of wavelet coefficients*, in preparation.
- [12] ———, *Fast wavelet techniques for near-optimal image processing*, in IEEE Military Communications Conference Record, San Diego, October 11–14, 1992, IEEE Press, Piscataway, NJ, 1992, pp. 1129–1135.
- [13] R. A. DeVore, P. Petrushev, and X. M. Yu, *Nonlinear wavelet approximations in the space $C(\mathbb{R}^d)$* , in Progress in Approximation Theory, Proceedings of the US/USSR Conference on Approximation, Tampa, 1990, Springer-Verlag, New York, 1992, pp. 261–283.
- [14] R. A. DeVore and V. Popov, *Interpolation of Besov spaces*, Trans. Amer. Math. Soc., 305 (1988), pp. 397–414.
- [15] R. A. DeVore and V. N. Temlyakov, *Some remarks on greedy algorithms*, Advances in Comp. Math., to appear.
- [16] D. Donoho, *De-noising by soft-thresholding*, IEEE Trans. Information Theory, 41 (1995), pp. 613–627.
- [17] ———, *Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition*, Appl. Comput. Harmonic Anal., 2 (1995), pp. 101–126.
- [18] D. Donoho and I. Johnstone, *Adapting to unknown smoothness via wavelet shrinkage*, J. Amer. Statist. Assoc., 90 (1995), pp. 1200–1224.
- [19] ———, *Ideal spatial adaptation by wavelet shrinkage*, Biometrika, 81 (1994), pp. 425–455.
- [20] ———, *Neo-classical minimax problems, thresholding and adaptive function estimation*, Bernoulli, 2 (1996), pp. 39–62.
- [21] D. Donoho, I. M. Johnstone, G. Kerkycharian, and D. Picard, *Wavelet shrinkage: Asymptopia?*, J. Roy. Statist. Soc. Ser. B, 57 (1995), pp. 301–369.
- [22] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North Holland, Amsterdam, 1976.
- [23] M. Frazier and B. Jawerth, *A discrete transform and decompositions of distribution spaces*, J. of Functional Anal., 93 (1990), pp. 34–170.
- [24] A. Harten, *Discrete multi-resolution analysis and generalized wavelets*, Appl. Numer. Math., 12 (1993), pp. 153–192, Special issue to honor Professor Saul Abarbanel on his sixtieth birthday.
- [25] C. Herley and M. Vetterli, *Biorthogonal bases of symmetric compactly supported wavelets*, in Wavelets, Fractals, and Fourier Transforms, Oxford Univ. Press, New York, 1993, pp. 91–1008.
- [26] D.-G. Kim, *Wavelet decompositions and function spaces on the unit cube*, Ph.D. Thesis, Purdue University, August 1994.
- [27] N.-Y. Lee and B. J. Lucier, *Inverting the Radon transform in the presence of noise*, in preparation.
- [28] B. J. Lucier, M. Kalelrgi, W. Qian, R. A. DeVore, R. A. Clark, E. B. Saff, and L. P. Clarke, *Wavelet compression and segmentation of mammographic images*, J. of Digital Imaging, 7 (1994), pp. 27–38.
- [29] P. Maass, *Wavelet-projection methods for inverse problems*, preprint.
- [30] Y. Meyer, *Ondelettes et Opérateurs I: Ondelettes*, Hermann, Paris, 1990; English transl. by D. H. Salinger, *Wavelets and Operators*, Cambridge Univ. Press, Cambridge, 1992.
- [31] J.-M. Morel and S. Solimini, *Variational Methods in Image Segmentation*, Progress in Nonlinear Differential Equations and Their Applications—Volume 14, Birkhauser, Boston, MA, 1994.
- [32] B. M. ter Haar Romeny, ed., *Geometry-driven Diffusion in Computer Vision*, Kluwer Acad. Publ., Dordrecht, the Netherlands, 1994.
- [33] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total varia-*

tion based noise removal algorithms, *Physica D*, 60 (1992), pp. 259–268.

- [34] G. Wahba, *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59, SIAM, Philadelphia, 1990.
- [35] R. White, *High-performance compression of astronomical images* (abstract only), in DCC 92: Data Compression Conference, J. Storer and M. Cohn, eds., IEEE Computer Society Press, Los Alamitos, CA, 1992, p. 403.