Wavelets, Approximation, and Compression

Martin Vetterli

Over the last decade or so, wavelets have had a growing impact on signal processing theory and practice, both because of their unifying role and their successes in applications (see also [42] and [38] in this issue). Filter banks, which lie at the heart of wavelet-based algorithms, have become standard signal processing operators, used routinely in applications ranging from compression to modems. The contributions of wavelets have often been in the subtle interplay between discrete-time and continuous-time signal processing. The purpose of this article is to look at recent wavelet advances from a signal processing perspective. In particular, approximation results are reviewed, and the implication on compression algorithms is discussed. New constructions and open problems are also addressed.

Bases, Approximation, and Compression

Finding a good basis to solve a problem dates at least back to Fourier and his investigation of the heat equation [18]. The series proposed by Fourier has several distinguishing features:

▲ The series is able to represent any finite energy function on an interval.

▲ The basis functions are eigenfunctions of linear shift invariant systems or, in other words, Fourier series diagonalize linear, shift invariant operators.

Similarly, the sinc basis used in the sampling theorem is able to represent any bandlimited function, and processing can be done on samples instead of the function itself. In short, a basis is chosen both for its ability to represent an object of interest (for example, good approximation with few coefficients) and for its operational value (for example, diagonalization of certain operators).

To be more formal, assume we have a space of functions S and we wish to represent an element $f \in S$. The space S can be, for example, integrable functions on the interval $[0,1]$ with finite square integral

$$\int_0^1 |f(t)|^2 \, dt < \infty, \tag{1}$$

which we denote $L_2(0,1)$. The first question is then to find a basis for S, that is a set of basis functions $\{\varphi_i\}_{i \in I}$ in S such that any element $f \in S$ can be written as a linear combination

$$f = \sum_{i \in I} \alpha_i \, \varphi_i. \tag{2}$$

The example closest to the heart of signal processing people is certainly the expansion of bandlimited functions in terms of the sinc function. Assume the space of real functions bandlimited to $[-\pi, \pi]$ in Fourier domain and having a finite square integral. We denote this space by $BL_2(-\pi, \pi)$. Then the Shannon sampling theorem says that any function in that space can be written as [36], [28]

$$f(t) = \sum_{n=-\infty}^{\infty} \alpha_n \, \mathrm{sinc}(t-n), \tag{3}$$

where $\alpha_n = f(n)$ are the samples of $f(t)$ at the integers, and

$$\mathrm{sinc}(t) = \frac{\sin(\pi t)}{\pi t} \tag{4}$$

and its integer shifts are the basis functions.

A question of immediate concern is in what sense (3) is actually true. Consider $\hat{f}_N(t)$ as the approximation using N terms, with N an odd integer

$$\hat{f}_N(t) = \sum_{n=-(N-1)/2}^{(N-1)/2} \alpha_n \mathrm{sinc}(t-n)$$

and a norm to measure the approximation error. In this paper, we are solely concerned with the L_2 norm of the error

$$\left\| f(t) - \hat{f}_N(t) \right\|_2 = \left(\int_{-\infty}^{\infty} \left| f(t) - \hat{f}_N(t) \right|^2 dt \right)^{1/2}. \tag{5}$$

Then we say that (3) is true in the sense that

$$\lim_{N \to \infty} \left\| f(t) - \hat{f}_N(t) \right\|_2 = 0. \tag{6}$$

If we have a set of functions $\{\varphi_i\}_{i \in I}$ such that all functions in S can be approximated arbitrarily precisely as in (6), then we say that the set is complete in S. If furthermore the elements are linearly independent, then we say that $\{\varphi_i\}_{i \in I}$ is a basis for S. Among all possible bases, orthogonal bases are particularly desirable. In that case, the basis functions or basis vectors are all mutually orthogonal. Normalizing their norm to 1, or $\|\varphi_i\|_2 = 1$, $i \in I$, an orthonormal set of vectors satisfies

$$\langle \varphi_i, \varphi_j \rangle = \delta_{ij}, \tag{7}$$

where the inner product $\langle ., . \rangle$ is defined as

$$\langle \varphi_i, \varphi_j \rangle = \int_{-\infty}^{\infty} \varphi_i^*(t) \varphi_j(t) dt \tag{8}$$

in continuous time, and as

$$\langle \varphi_i, \varphi_j \rangle = \sum_{n \in \mathbb{Z}} \varphi_i^*[n] \varphi_j[n] \tag{9}$$

in discrete time. (For discrete-time signals, we will consider the space of square summable sequences denoted by $l_2(Z)$.) For orthonormal bases, the expansion formula (2) becomes

$$f = \sum_{i \in I} \langle \varphi_i, f \rangle \varphi_i. \tag{10}$$

This is a representation formula for f in the orthonormal basis $\{\varphi_i\}_{i \in I}$. (More general formulas, for example in biorthogonal bases or in frames, are also possible, but are not needed for our discussion.) Given the representation of f in an orthonormal basis as in (10), its orthogonal projection into a fixed subspace of dimension N spanned by $\{\varphi_n\}_{n=0...N-1}$ is denoted $\hat{f}_N(t)$

$$\hat{f}_N = \sum_{n=0}^{N-1} \langle \varphi_n, f \rangle \varphi_n. \tag{11}$$

This is the linear approximation of f by a projection onto a fixed subspace of dimension N. It is linear because the approximation of $\alpha f + \beta g$ is equal to the weighted sum of approximations, $\alpha \hat{f} + \beta \hat{g}$, as can be readily verified. But this is only one of many possible N-term approximations of f. Thus, we are led to the approximation problem: Given objects of interest and spaces in which they are embedded, we wish to know how fast an N-term approximation converges

$$\left\| f(t) - f_N(t) \right\|_2 \sim \text{fct}(N), \tag{12}$$

where $f_N(t)$ stands for an approximation of $f(t)$ which involves N elements, to be chosen appropriately.

This immediately raises a number of questions. Different bases can give very different rates of approximation. Then, there are various ways to choose the N terms used in the approximation. A fixed subset (e.g., the first N terms)

If a function is piecewise smooth, with isolated discontinuities, then Fourier approximation is poor because of the discontinuities.

leads to a linear, subspace approximation as in (11). Adaptive schemes, to be discussed later, are nonlinear. Will different choices of the subset lead to different rates of approximation? Such questions are at the heart of approximation theory and are relevant when choosing a basis and an approximation method for a given signal processing problem. For example, denoising in wavelet bases has led to interesting results for piecewise smooth signals precisely because of the superior approximation properties of wavelets for such signals.

We are now ready to address the last problem we shall consider, namely the compression problem. This involves not only approximation quality, but also description complexity. There is a cost associated with describing f_N, and this cost depends on the approximation method. Typically, the coefficient values and their locations need to be described, which involves quantization of the coefficients and indexing their locations.

Calling R the number of bits used in the description, we can define a function $\hat{D}(R)$ as

$$\hat{D}(R) = \| f - f_R \|_2, \tag{13}$$

where f_R explicitly indicates that R bits are used to approximate f. (At this point, we should introduce an expectation over the class of functions to approximate. Formally, approximating a single function has zero complexity in the information theoretic sense. For simplicity, we leave this implicit for the moment.) Of course, f_R depends on the basis chosen to represent f, and thus $\hat{D}(R)$ can have very different behaviors, depending on the basis. While $\hat{D}(R)$ resembles the distortion rate function $D(R)$ defined in information theory, the two functions are not identical by any means. $D(R)$ is defined as the infimum of the distortion of any coding scheme using R bits [7] (using arbitrary long blocks of samples in discrete time, for example) while $\hat{D}(R)$ is geared specifically at approximation in bases. In certain cases, $\hat{D}(R)$ can be of the same order as $D(R)$ while in others, the theoretical minimum given by $D(R)$ can be much better. Yet, $\hat{D}(R)$ is usually achievable in practice with reasonable complexity, while $D(R)$ often remains an unachievable limit. (It should also be noted that the true distortion-rate function is known only in very few cases.)

To wrap up this section, we reviewed three related problems around bases in spaces. The first was the existence of a basis, the second was the approximation power of a basis, and the third was the compression power. In the

sequel, we shall explore the relationship of these problems, in particular for wavelet bases. The review paper [15] also discusses the relationship between harmonic analysis, computational methods, and signal compression.

Approximation Properties of Filter Banks and Wavelets

Although orthogonal filter banks as used in subband coding have been discussed by Usevitch [42], we review them here with an emphasis on their projection and approximation properties.

Orthogonal Filter Banks

When thinking of filtering, one usually thinks about frequency selectivity. For example, an ideal discrete-time lowpass filter with cut-off frequency $\omega_c < \pi$ takes any input signal and projects it onto the subspace of signals bandlimited to $[-\omega_c, \omega_c]$. Orthogonal discrete-time filter banks perform a similar projection which we now review. Assume a discrete-time filter with finite impulse response $g_0[n] = \{g_0[0], g_0[1], \ldots, g_0[L-1]\}$, L even, and the property

$$\langle g_0[n], g_0[n-2k] \rangle = \delta_k, \tag{14}$$

that is, the impulse response is orthogonal to its even shifts, and $\|g_0\|_2 = 1$. Denote by $G_0(z)$ the z-transform of the impulse response $g_0[n]$

$$G_0(z) = \sum_{n=0}^{L-1} g_0[n] z^{-n}, \tag{15}$$

with an associated region of convergence covering the z-plane except the origin. Assume further that $g_0[n]$ is a lowpass filter, that is, its discrete-time Fourier transform has most of its energy in the region $[-\pi/2, \pi/2]$. Then define a high-pass filter $g_1[n]$ with z-transform $G_1(z)$ as follows:

$$G_1(z) = z^{-L+1} G_0(-z^{-1}). \tag{16}$$

Three operations have been applied:
▲ $z \rightarrow -z$ corresponds to modulation by $(-1)^n$, or transforming the lowpass into a highpass.
▲ $-z \rightarrow -z^{-1}$ applies time-reversal to the impulse response.
▲ Multiplication by z^{-L+1} makes the time-reversed impulse response causal.
For example, if $g_0[n] = \{a, b, c, d\}$, then $g_1[n] = \{-d, c, -b, a\}$. This special way of obtaining a highpass from a lowpass, introduced as conjugate quadrature filters (QCF) [39], [27], has the following properties:

$$\langle g_1[n], g_1[n-2k] \rangle = \delta_k, \tag{17}$$

that is, $g_1[n]$ is also orthogonal to its even shifts, and

$$\langle g_0[n], g_1[n-2k] \rangle = 0 \tag{18}$$

or $\{g_0[n], g_1[n]\}$ and their even shifts are mutually orthogonal. The last step is to show that the orthonormal set

$$\{g_0[n-2k], g_1[n-2l]\}_{k,l \in \mathbb{Z}} \tag{19}$$

is an orthonormal basis for $l_2(\mathbb{Z})$, the space of square summable sequences. This requires to check completeness, which can be done for example by verifying Parseval's relation [44]. Thus, any sequence from $l_2(\mathbb{Z})$ can be written as

$$x[n] = \sum_{k \in \mathbb{Z}} \alpha_k \cdot g_0[n-2k] + \sum_{l \in \mathbb{Z}} \beta_l \cdot g_1[n-2l], \tag{20}$$

where

$$\alpha_k = \langle g_0[n-2k], x[n] \rangle, \; k \in \mathbb{Z} \tag{21}$$

$$\beta_l = \langle g_1[n-2l], x[n] \rangle, \; l \in \mathbb{Z}. \tag{22}$$

Now, how do we implement the expansion formula (20)? Simply using a multirate filter bank as depicted in Fig. 1. The analysis filters $H_0(z)$ and $H_1(z)$ are time-reversed versions of $G_0(z)$ and $G_1(z)$, or

$$H_i(z) = \widetilde{G}_i(z) = G_i(z^{-1}), \; i = 0,1. \tag{23}$$

It is not hard to verify that the analysis filter bank (the left part of Fig. 1) computes indeed the coefficients α_k and β_l (remember that convolution includes time reversal) and that the synthesis filter bank (the right part of Fig. 1) realizes the reconstruction as in (20).

We shall have a second look at the lowpass analysis/synthesis branch, as shown in Fig. 2. Using standard multirate signal processing analysis, the output of such a filtering-downsampling-upsampling-filtering gives an output $y_0[n]$ with z-transform

$$Y_0(z) = \frac{1}{2} \cdot G_0(z) \cdot [H_0(z)X(z) + H_0(-z)X(-z)], \tag{24}$$

that is, both a filtered version of $X(z)$ and a filtered aliased version involving $X(-z)$. It can be verified that the operator from $x[n]$ to $y_0[n]$ is linear, self-adjoint, and idempotent, that is, it is an orthogonal projection of the input space $l_2(\mathbb{Z})$ onto a subspace given by the span of $\{g_0[n-2k]\}_{k \in \mathbb{Z}}$. We call this subspace V_1, for which $\{g_0[n-2k]\}_{k \in \mathbb{Z}}$ forms an orthonormal basis.

Considering similarly the highpass branch, we see that it produces a projection of $l_2(\mathbb{Z})$ onto a subspace W_1 given by the span of $\{g_1[n-2l]\}_{l \in \mathbb{Z}}$. The spaces V_1 and W_1 are orthogonal (see (18)) and their direct sum forms $l_2(\mathbb{Z})$, or

$$l_2(\mathbb{Z}) = V_1 + W_1. \tag{25}$$

Thus, an orthogonal filter bank splits the input space into a lowpass approximation space V_1 and its (highpass) or-
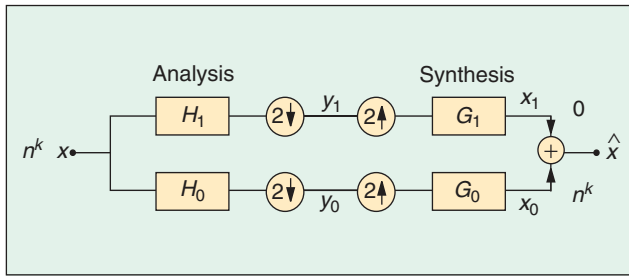
▲ 1. Multirate filter bank. Two channel analysis followed by synthesis.
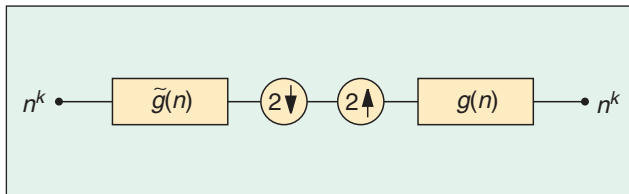
▲ 2. Lowpass branch of the two channel filter bank in Fig. 1. The discrete-time polynomial n^k is an eigensignal of this operator.

thogonal complement W_1. The space V_1 corresponds to a coarse approximation, while W_1 contains additional details. This is the first step in the multiresolution analysis that is obtained when iterating the highpass/lowpass division on the lowpass branch (see Fig. 3).

Discrete-Time Polynomials and Filter Banks

Signal processing specialists intuitively think of problems in terms of sinusoidal bases. Approximation theory specialists think often in terms of other series, like the Taylor series, and thus, of polynomials as basic building blocks. We now look at how polynomials are processed by filter banks.

A discrete-time polynomial signal of degree M is composed of a linear combination of monomial signals

$$p^{(m)}[n] = n^m \quad 0 \le m \le M. \qquad (26)$$

We shall now see that such monomial (and therefore polynomial) signals are eigensignals of certain multirate operators. We need to consider lowpass filters $G_0(z)$ that have a certain number $N > 0$ of zeroes at $z = -1$, or $\omega = \pi$ on the unit circle. That is, the filter factors as

$$G_0(z) = (1 + z^{-1})^N R_0(z). \qquad (27)$$

Clearly, because of (16), the highpass $G_1(z)$ has N zeros at $z = 1$ (or $\omega = 0$), while $H_0(z)$ and $H_1(z)$ have N zeros at $z = -1$ and 1, respectively because of (23).

Then, monomials up to degree $N - 1$ are eigensignals of the lowpass branch shown in Fig. 2, that is, filtering by $H_0(z) = G_0(z^{-1})$, downsampling by 2, upsampling by 2, and finally filtering by $G_0(z)$. This result is a consequence of the Strang-Fix theorem (see for example [41]), and we give here an intuitive reasoning. Because the highpass branch has an analysis filter

$$H_1(z) = z^{L-1}G_0(-z) = z^{L-1}(1 - z^{-1}) \cdot R_0(-z), \qquad (28)$$

with N zeros at $z = 1$, all polynomials up to degree $N - 1$ get zeroed out in this branch. This is a consequence of the moment theorem of the Fourier transform [44], [28]. Since the filter bank has the perfect reconstruction property, these polynomials are therefore perfectly replicated by the lowpass branch only, showing that they are eigensignals. It is worth pointing out that sinusoids are not eigensignals of such multirate operators. A consequence of this eigensignal property is that we can write

$$n^m = \sum_{k=-\infty}^{\infty} \alpha_k^{(m)} \cdot g_0[n - 2k] \quad m \le N, \qquad (29)$$

where $\alpha_k^{(m)}$ are appropriate coefficients. This shows that monomials up to degree $N - 1$ are in the span of $\{g_0[n - 2k]\}_{k \in \mathbb{Z}}$. (Note that polynomials are not in $l_2(\mathbb{Z})$, but any restriction to an interval is. So (29) is true for windowed polynomials up to boundary effects.) Consider now an iterated filter bank as shown in Fig. 3. After J stages of decomposition and synthesis, the equivalent synthesis lowpass filter $g_0^{(J)}[n]$ has z-transform

$$G_0^{(J)}(z) = \prod_{j=0}^{J-1} G_0(z^{2^j}), \qquad (30)$$

as can be verified using standard multirate identities [43], [44].

Consider now what happens if a monomial signal of degree smaller than N as given in (27) enters the analysis filter bank. It gets cancelled in all highpass filters and gets thus reproduced by the lowpass branch alone. In other words, linear combinations of $g_0^{(J)}[n]$ and its shifts by 2^J can reproduce polynomials up to degree $N - 1$, or similarly to (29), we can write

$$n^m = \sum_l \alpha_{J,l}^{(m)} \cdot g_0^{(J)}[n - 2^J l]. \qquad (31)$$

Therefore, we have seen that discrete-time polynomials live in the "coarse" approximation space of discrete-time filter banks, and this up to degree $N - 1$ when the lowpass filter has N zeros at $z = -1$.

Continuous-Time Polynomials and Wavelets

As is well known, a strong link exists between iterated filter banks and wavelets. For example, filter banks can be used to generate wavelet bases [9], and filter banks can be used to calculate wavelet series [23]. It comes thus as no surprise that the properties seen in discrete time regarding polynomial representation carry over to continuous time. While these properties are directly related to moment properties of wavelets and thus hold in general, we review them in the context of wavelets generated from orthogonal finite impulse response (FIR) filter banks. Assume again that the lowpass filter has N zeros at $\omega = \pi$, and thus, the highpass has N zeros at $\omega = 0$. From the two

scale relation of scaling function and wavelet, we get that the Fourier transform of the wavelet can be factored as

$$\Psi(\omega) = \frac{1}{\sqrt{2}} G_1 \left(e^{j\omega/2} \right) \cdot \Phi\left(\frac{\omega}{2}\right), \tag{32}$$

where $\Phi(\omega)$ is the Fourier transform of the scaling function. Since $\Phi(0) = 1$, we see from (32) that $\Psi(\omega)$ has exactly N zeros at $\omega = 0$, or N zero moments following the moment theorem of the Fourier transform. Therefore, the inner product between any wavelet (arbitrary scale and shift) with polynomials up to degree $N - 1$ will be zero. Thus, polynomials can be represented as linear combinations of scaling functions alone, or similarly to (31), there are coefficients $\gamma_l^{(m)}$ such that

$$t^m = \sum_l \gamma_l^{(m)} \varphi(t - l). \tag{33}$$

An example, using the Daubechies scaling function based on a four-tap filter [9], is shown in Fig. 4 over a finite interval. The linear function is perfectly represented, up to boundary effects.

Discontinuities in Filter Bank and Wavelet Representations

What happens if a signal is discontinuous at some point t_0? We know that Fourier series do not like discontinuities, since they destroy uniform convergence. Wavelets have two desirable properties as far as discontinuities are concerned. First, they focus locally on the discontinuity as we go to finer and finer scales. That is because of the scaling relation of wavelets, where the function set $\psi_{m,n}(t)$ is defined as

$$\psi_{m,n}(t) = 2^{-m/2} \, \psi(2^{-m} t - n) \quad m, n \in \mathbb{Z}, \tag{34}$$

where $m \to -\infty$ corresponds to fine details. Thus, as m grows negative, the wavelet becomes "sharper." If the discontinuity is isolated, and the surroundings are smooth, all wavelet inner products except the ones at the discontinuity will be zero, and around the discontinuity, $L - 1$ inner products are different from zero when the wavelet has support length L.

Second, the magnitude evolution across scales of the nonzero wavelet inner products characterizes the discontinuity. This is a well-known characteristic of the continuous wavelet transform [25], [10] and holds as well for the

orthonormal wavelet series. For example, for a Dirac, $f(t) = \delta(t)$, the wavelet coefficients, given by

$$c_{m,n} = \langle \psi_{m,n}, f \rangle,$$

will evolve as

$$|c_{m,n}| \sim 2^{-m/2}, \tag{35}$$

for n around zero, as can be seen from (34). For a Heaviside (step) function the behavior is

$$|c_{m,n}| \sim 2^{m/2}, \tag{36}$$

and more generally, for a singularity of order k (a singularity of order k appears when there are k bounded derivatives, and the $(k+1)$th derivative is unbounded. The Heaviside function has $k = 0$, a continuous function with a nondifferentiable point has $k = 1$), the behavior is

$$|c_{m,n}| \sim 2^{(k+\frac{1}{2}) m}, \tag{37}$$

for the $L - 1$ coefficients around the discontinuity at scale m. Consider now a signal that is mostly smooth, except at isolated points where it has various types of singularities. Clearly, the energy of the wavelet coefficients will be concentrated around the points of singularity, with a behavior indicated by (37). If the signal is polynomial between the singularities, then all other wavelet coefficients will be zero, and the polynomial pieces will be "caught" by the scaling functions. This is shown in Fig. 5.

The above discussion focused on continuous-time signals, so what about discrete-time one? Technically, there is no concept of continuity for sequences. However, qualitatively, the same happens as in continuous time. In addition, the sequence can be piecewise polynomial, in which case we have a piecewise equivalent to the above discussion. In that case, the lowpass branch catches the coarse trend, while the highpass and bandpass channels are zero except around singular points. At these points, there will be $L - 1$ nonzero coefficients at each scale, and their magnitude will be governed as in (37).

To conclude, we have discussed how piecewise polynomials or piecewise smooth functions have wavelet expansion coefficients concentrated mostly around singular
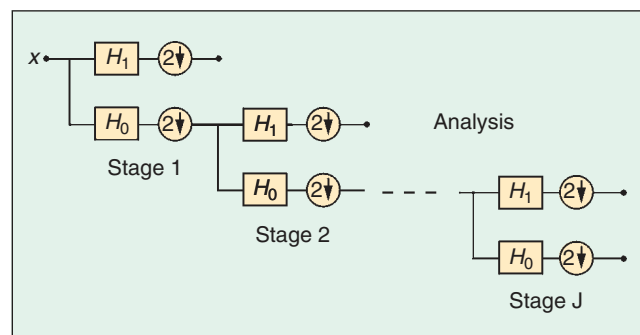
▲ 3. Iterated filter bank. The lowpass branch gets split repeatedly to get a discrete-time wavelet transform.

points, while scaling functions or lowpass filters describe the polynomial or smooth trend.

## Linear and NonLinear Approximations

So far, we have reviewed the construction of bases, especially of wavelet bases. In that case, we were concerned with an exact representation of any signal or function from a given space. We turn now to approximate representations and contrast linear with nonlinear methods. For an in-depth discussion of such methods, in particular in the wavelet context, we refer to the excellent exposition in [24].

Linear Approximation
Assume a space V and an orthonormal basis $\{g_n\}_{n \in N}$ for V. Thus, a function $f \in V$ can be written as a linear combination

$$f = \sum_{n \in N} \langle g_n, f \rangle g_n. \tag{38}$$

The best (in the squared error sense) linear approximation of f in the subspace V, denoted as $\hat{f}$, is given by the orthogonal projection of f onto a fixed subspace of V (see
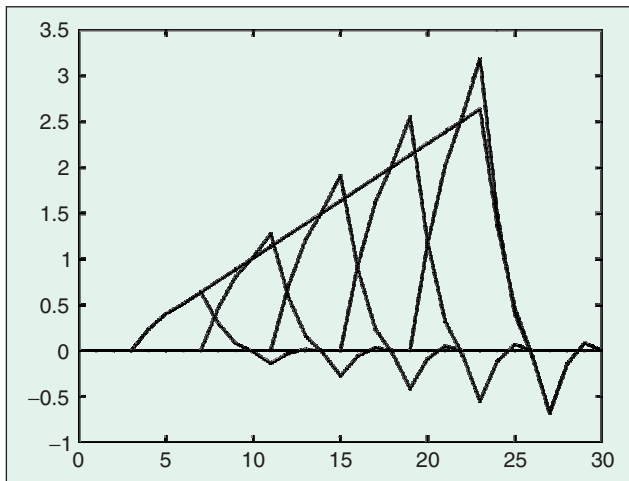
▲ 4. *Reproduction of a linear function as a combination of scaling functions.*
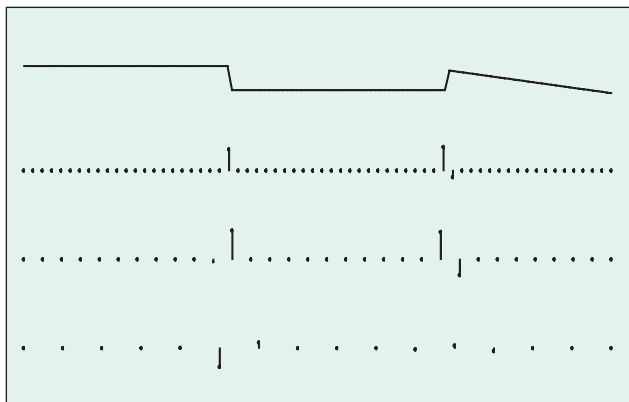
▲ 5. *A piecewise smooth function and its wavelet expansion. The wavelet coefficients are different from zero only in the vicinity of the discontinuity, while the coarse behavior is represented by the scaling coefficients only.*

(11)). Assume it is a subspace W of dimension M, and spanned by the first M vectors of the basis, or $W = span\{g_0, g_1, g_2, ..., g_{M-1}\}$. Then $\hat{f}$ is given by

$$\hat{f}_M = \sum_{n=0}^{M-1} \langle g_n, f \rangle g_n, \tag{39}$$

and the squared error of the approximation is

$$\hat{\varepsilon} = \left\| f - \hat{f} \right\|_2^2 = \sum_{n=M}^{\infty} \left| \langle g_n, f \rangle \right|^2. $$

Because the subspace W is fixed, independent of f, the approximation is linear. This follows from the fact that the approximation of $f + g$ equals the sum of their approximations. An obvious question is to find the best basis for linear approximation (in the squared error sense) given a class of objects. To be specific, we review the classic result on linear approximation of random vector processes. Consider a vector process $X = [X_0, X_1, X_2, ..., X_{N-1}]^T$ where X is an independent, identically distributed (iid) process with zero mean ($E[X_i] = 0$) and covariance

$$E[X \cdot X^T] = R_X, \tag{40}$$

where R_X is known a priori. Because we are interested in least squares approximation, we will not need any higher order statistical characterization. Consider now the linear approximation in an orthonormal basis $\{g_0, g_1, ..., g_{M-1}\}$, $M < N$, that is

$$\hat{X}_M = \sum_{n=0}^{M-1} \langle g_n, X \rangle g_n. \tag{41}$$

What is the expected squared error, or $\hat{\varepsilon}_M = E\left[\left\| X - \hat{X}_M \right\|^2 \right]$? Rewriting $\left\| X - \hat{X}_M \right\|$ as $\langle (X - \hat{X}_M), (X - \hat{X}_M) \rangle$, one finds after some manipulation that [20]

$$\left\| X - \hat{X} \right\|^2 = \sum_{m=M}^{N-1} g_m^T X X^T g_m, \tag{42}$$

which leads to the expected squared error

$$\hat{\varepsilon}_M = \sum_{m=M}^{N-1} g_m^T E[X X^T] g_m = \sum_{m=M}^{N-1} g_m^T \cdot R_X \cdot g_m. \tag{43}$$

The question is now: over all possible orthonormal bases for $\mathbb{R}^N$, which will minimize (43) for all $M = 1, ..., N-1$? Given that R_X is positive semidefinite, the answer is that the best basis is given by the set of eigenvectors of R_X ordered with decreasing eigenvalues. The approximation error is then

$$\hat{\varepsilon}_M = \sum_{m=M}^{N-1} \lambda_m, \tag{44}$$

where λ_m is the mth largest eigenvalue of R_X. The geometric intuition behind this result is that the eigenvectors are the principal axes of the N-dimensional distribution and the best M-dimensional subspace gathering most "energy" (or expected l_2 norm) corresponds to the principal axes with largest eigenvalues (see Fig. 6).

The orthogonal transform given by the eigenvectors of R_X (ordered by decreasing eigenvalues) is of course the well-known Karhunen-Loève transform (KLT) as explained by Goyal [20], but it is useful to recapitulate its essential property: it is the best basis for linear subspace approximation of a vector process specified by second order statistics. Again, the second-order statistics are assumed to be known a priori, so the subspace is chosen a priori as well and is fixed. (If the correlation matrix has to be estimated on the fly, then we have a different, adaptive situation, which is not a linear approximation. For an example of such an adaptive KLT, see [22].)

In the case of a jointly Gaussian vector process, a stronger result holds, namely, it is also the best "compression" transform (when transform coefficients are quantized and entropy coded). See [19]-[21] for a review of this result.

Nonlinear Approximation in Orthonormal Bases

Consider the same set up as above, but with a different approximation rule. Instead of (38), where the first M coefficients in the orthonormal expansion are used, we keep the M largest coefficients instead. That is, we define an index set I_M of the M largest inner products, or

$$\left| \langle g_m, f \rangle \right| \geq \left| \langle g_n, f \rangle \right| \text{ for every } m \in I_m \text{ and } n \notin I_m. \quad (45)$$

Then, we define, the best nonlinear approximation as:

$$\widetilde{f}_M = \sum_{n \in I_M} <g_n, f> g_n, \quad (46)$$

which leads to an approximation error

$$\widetilde{\varepsilon}_M = \left\| f - \widetilde{f} \right\|_2^2 = \sum_{n \notin I_M} |< g_n, f >|^2.$$

Clearly

$$\widetilde{\varepsilon}_M \leq \hat{\varepsilon}_M. \quad (47)$$

(We could call this adaptive linear approximation or adaptive subspace approximation. However, the commonly used term is nonlinear approximation. More general nonlinear schemes could also be considered, but are beyond the scope of this article.)

Let us show that the simple modification of picking the largest rather than the first M inner products creates indeed a nonlinear scheme. Consider two functions f and h. In general, the M largest coefficients will not correspond to the same set for f and h, and thus, approximat-
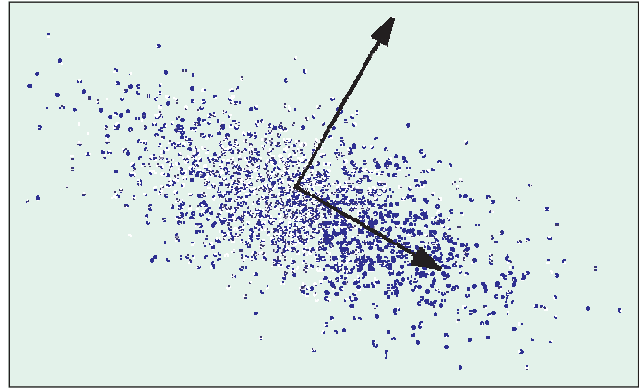
▲ 6. Two-dimensional distribution indicated by a scatter plot. The principal axes are indicated, and the best one-dimensional approximation is the axis corresponding to the biggest spread.

ing $f + h$ is certainly not the sum of the approximations of f and h, which has typically more than M terms.

Linear versus Nonlinear Approximation

Now, the question is if (45) leads to large differences in approximation quality, and if there are bases where the difference is larger than in others. Consider the example of a piecewise constant function with a discontinuity. While simple, it exposes fundamental differences between linear and nonlinear approximation, as well as between Fourier and wavelet bases.

To be specific, take the interval $[0,1]$ and use Fourier series or wavelet series to represent a function which is constant except for a jump at a random point t_0 distributed uniformly over $[0,1]$.

Consider Fourier series first. Because of the discontinuity, the Fourier series coefficients decay slowly

$$\left| \alpha_n \right|, \left| \beta_n \right| \sim 1/n, \quad (48)$$

where α_n and β_n are the cosine and sine coefficients at frequency $2\pi n$. Linear approximation by the first M terms (or first $M/2$ sines and cosines) leads to an error of the order

$$\hat{\varepsilon}_M^{(F)} \sim \sum_{n=M}^{\infty} 1/n^2 \sim \frac{1}{M}. \quad (49)$$

Picking the M largest coefficients will not change the approximation order for the following reason. While some of the α_n, β_n coefficients will be small in the early terms of expansion (the sine and cosine in the numerator of the Fourier series coefficient can sometimes be small), most of the M largest coefficients are still gathered in the first kM coefficients, where k is a small integer. So, to first approximation, the set of M largest coefficients has a large overlap with the set of the first M coefficients, and thus the nonlinear approximation error satisfies

$$\widetilde{\varepsilon}_M^{(F)} \sim \frac{1}{M}. \quad (50)$$

Consider now the wavelet expansion using a Haar wavelet, where

$$\Psi(t) = \begin{cases} 1 & 0 \leq t < 1/2 \\ -1 & 1/2 \leq t < 1 \\ 0 & \text{else} \end{cases}$$

(51)

and the wavelets are defined as in (34) with $m = 0, -1, -2, \ldots$ While such a wavelet seems unfairly close to the function to be represented, let us point out that there are still an infinite number of nonzero wavelet coefficients unless t_0 is a dyadic rational. In addition, the behavior we are about to show holds much more generally.
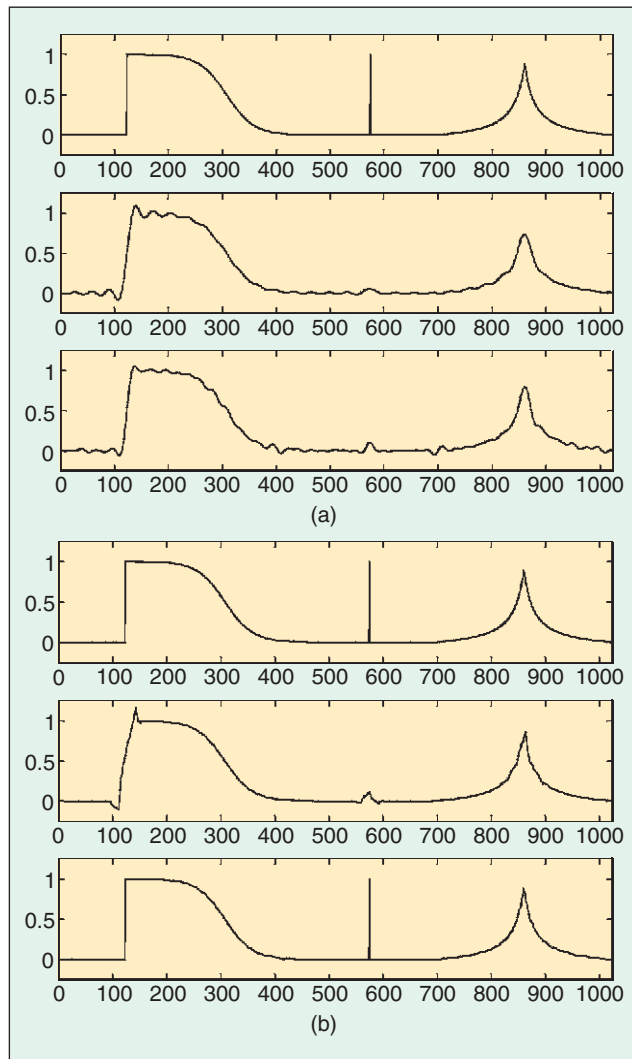
▲ 7. Fourier versus wavelet bases and linear versus nonlinear approximation. The signal is discrete time of length 1024, and $M = 64$ coefficients are retained. (a) Fourier case. Original on top, linear approximation in the middle, and nonlinear approximation at the bottom. The MSE is 2.7 and 2.4 respectively. (b) Wavelet case, with six levels of decomposition and a wavelet having three zero moments. Original on top, linear approximation in the middle, and nonlinear approximation at the bottom. The MSE is 3.5 and 0.01, respectively.

First, recall that there will be a single nonzero wavelet coefficient at each scale m around the discontinuity, and this coefficient has size

$$|c_{m,n}| \sim 2^{m/2},$$

(52)

where $m \to -\infty$. When doing linear approximation, one takes coarser scales first (where the larger coefficients lie), and goes to finer scales next. While all coefficients but one are zero, we are not allowed to adapt. If $M = 2^J$, one can take $J - 1$ scales (from 0 to $-J + 2$), at which point the only nonzero coefficient is of order $2^{-J/2}$. The residual quadratic error is then of order 2^{-J} and thus

$$\hat{\varepsilon}_M^{(W)} \sim \frac{1}{M},$$

(53)

just as in the Fourier case. However, if we go to the nonlinear approximation scheme, things change dramatically. Picking the M largest coefficients allows us to go across M scales (since there is only one nonzero coefficient per scale) at which point the coefficient is of order $2^{-M/2}$. The quadratic error is then

$$\widetilde{e}_M^{(W)} \sim 2^{-M}.$$

(54)

We have gone from $1/M$ decay in Fourier (linear and nonlinear) or wavelet linear approximation, to exponential decay with nonlinear approximation in wavelet bases.

What we have just seen holds much more generally. If a function is piecewise smooth, with isolated discontinuities, then Fourier approximation is poor because of the discontinuities. In the wavelet case with a wavelet that has enough zero moments so that the inner products in the smooth region are small, few wavelet coefficients are sufficient to capture the discontinuities, and nonlinear approximation outperforms linear schemes by orders of magnitude. Fig. 7 shows an example, where it can be seen that nonlinear approximation is vastly superior in the wavelet case, while it makes little difference in the Fourier case. Also, note that with linear approximation, there is little difference between Fourier and wavelets. For a review of some strong approximation results for wavelet expansions of smooth and piecewise smooth functions, see [12], [24].

Nonlinear Approximation and Compression

So far, we have considered keeping M elements from a basis, either a fixed set (the first M typically) or an adaptive set (corresponding to the largest projection). In compression, we have first to describe the coefficient set, which has zero cost when this set is fixed (linear approximation) but has a nontrivial cost when it is adaptive (nonlinear approximation). In that case, there are $\binom{N}{M}$ possible subsets, and the rate to describe the subset is equal to the entropy of the distribution of the subsets, or at most

$$\log_2 \binom{N}{M}. \tag{55}$$

In addition, it is necessary to quantize and entropy code the various inner products. For complexity reasons, one usually considers only scalar quantization and entropy coding of individual coefficients, rather than vector quantization and vector entropy coding. Using usual high rate quantization results [19], [20], spending R bits on a particular coefficient leads to a distortion-rate trade-off of the form

$$D(R) = C_d \, 2^{-2R}, \tag{56}$$

where C_d depends on the distribution of the coefficient (e.g., σ^2 for a uniform distribution of variance σ^2 or $(\pi e/6)\sigma^2$ for a Gaussian random variable of variance σ^2). Using the M-term approximation rates (as for example in (49), (50) or (53), (54)) together with the set description complexity (55) and distortion-rate for coefficients (56), one can derive achievable upper bounds for the distortion-rate tradeoff of linear and nonlinear approximation schemes.

An Example of Linear versus Nonlinear Approximation and Compression

An example that highlights the fundamental difference between linear and nonlinear approximation in compression is as follows. Consider a random vector process of size N, where each realization has only one nonzero value at a random location k, the others being zero, or

$$X[n] = \alpha \cdot \delta[n-k] \quad n \in \{0,\dots,N-1\}, \tag{57}$$

where α is Gaussian $N(0,\sigma^2)$ and k is an integer uniformly distributed over $\{0\dots N-1\}$. In other words, for each realization, we pick uniformly a location between 0 and $N-1$, and at that location, place a Gaussian random variable of mean zero and covariance σ^2, as shown schematically in Fig. 8.

Consider first the linear approximation problem. The autocovariance matrix is equal to

$$R_X = \frac{\sigma^2}{N} \cdot I, \tag{58}$$

and therefore, using the KLT approach, the standard basis is already the best basis (not unique) for linear subspace approximation. Thus, R bits are evenly distributed across N positions, and the overall distortion-rate behavior is

$$D_L(R) = C \cdot \sigma^2 \cdot 2^{-2R/N}. \tag{59}$$

Note that this is what the linear, KLT approach would tell us to do, even though it is clearly suboptimal. A better approach, at least if R is large enough, is to spend $\lceil \log_2 N \rceil$ bits to point to the location that is active in a given realization, and spend the remaining $R - \log_2 N$

$R - \log_2 N$ bits on describing the Gaussian random variable. This leads to

$$D_{NL}(R) = C \cdot \sigma^2 \cdot 2^{-2(R-\log_2 N)}, \tag{60}$$

which for $R \gg \log_2 N$ is clearly superior to (59). For low rates ($R \sim \log_2 N$ and below), the situation is more difficult; refer to Weidmann [45] for a thorough analysis.

The lesson from the above example is that the linear theory is misleading when processes are far from jointly Gaussian. The KLT leads to decorrelation, which means independence only in the Gaussian case. And it turns out that many signals of interest for compression are not jointly Gaussian.

### Real Coders Do Nonlinear Approximation

Let us see what "real" coders do. Actually, long before wavelets and nonlinear approximation theory, compression engineers developed methods to adapt the linear, KLT-based theory to real world signals. Two main ingredients are used, namely locality and adaptivity. First, signals are divided into localized subsignals (for example using blocking as in block-based transform coding), and these subsignals are transformed individually. Then, these "local" transforms are coded adaptively, depending on their characteristics. This is reminiscent of our previous example: if only a few large coefficients appear but in unknown locations, then it pays to localize them. Without getting into too much details, Fig. 9 shows a typical set of discrete cosine transform (DCT) coefficients as used in JPEG and the subsequent quantization, run length, and entropy coding. With this method, when there are few large coefficients, they can be represented with few bits, and we have indeed a nonlinear approximation scheme in a local Fourier-like basis.

It is worthwhile to point out that any adaptive scheme creates nonlinearity. For example, adaptive bit allocation between transform blocks (more bits to high activity blocks, less to low activity ones) is already a nonlinear approach. Also, adaptive best bases [6], [34], dynamic programming based allocations [29], [34], adaptive segmentation [31], [32], and matching pursuit [26] are all examples of nonlinear schemes used in compression.

It is worth pointing out that the investigation of the rate-distortion behavior of "real" coders is a difficult topic, both because of the nonlinearity of approximation
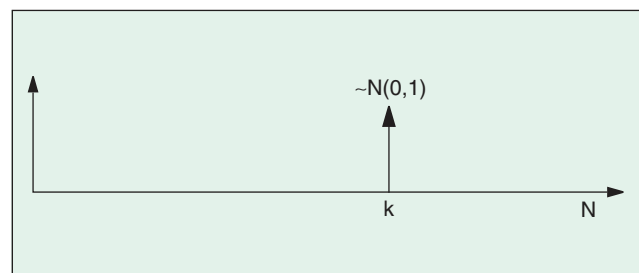
▲ 8. Pointing process, with uniform choice of a single location where a Gaussian random variable is placed.
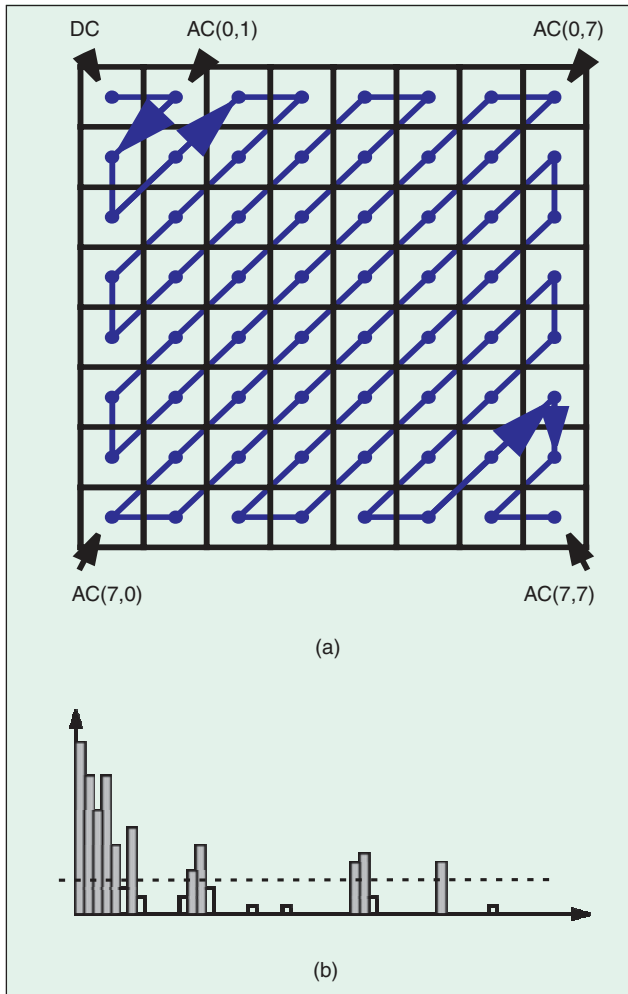
▲ 9. Example of 64 DCT coefficients from an 8 × 8 block and zig-zag scanning. (a) Zig-zag scanning of the 8 × 8 block of DCT coefficients to obtain a one-dimensional array of coefficients. (b) Example of an array of 64 coefficients (magnitude only) with a threshold T (deadzone quantizer) and quantization of the coefficients above T. Subsequently, run length coding is used to index the large coefficients, and entropy coding to represent their value.
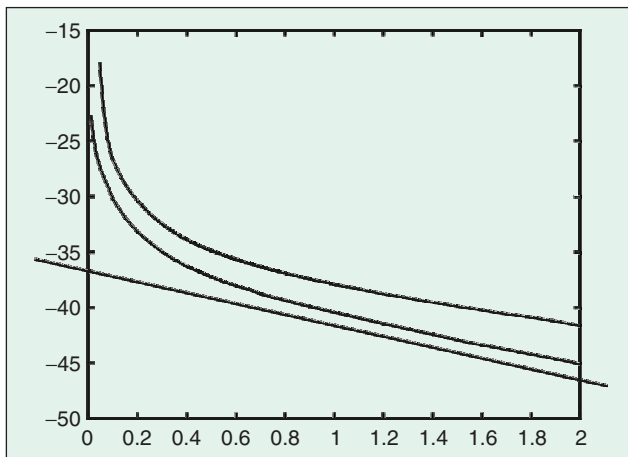
▲ 10. Squared error (on log scale) versus bit rate for JPEG (upper curve) and SPIHT [35] (lower curve). At high rates, the typical −6 dB/bit is apparent, while at very low rates, a steeper decay is typical.

and quantization and because the interplay of approximation, quantization and entropy coding is nontrivial. In particular, the low rate behavior is tricky. Fig. 10 shows the logarithm of the squared error of JPEG and of a wavelet-based coder (SPIHT [35]). While the high rate behavior is as expected (namely, −6 dB per bit on log scale), the low rate behavior is different, since the quadratic error falls off much more rapidly at very low rates. An intuitive explanation is that at very low rates, the classic rate allocation does not hold, since many transform coefficients do not get any rate at all. The few remaining coefficients are thus getting more precision, leading to a faster decay of the error. Using a different analysis, Mallat et al. [24] have shown a $1/R$ behavior at low rates for transform codes.

Compression of Piecewise Polynomial Signals

Let us return to one-dimensional piecewise smooth signals. Wavelets are well suited to approximate such signals when nonlinear approximation is allowed. To study compression behavior, consider the simpler case of piecewise polynomials, with discontinuities. To make matters easy, let us look again at the signal we used earlier to study non-linear approximation, but this time include quantization and bit allocation. A simple analysis of the approximate rate distortion behavior of a step function goes as follows. Coefficients decay as $2^{m/2}$, so the number of scales J involved, if a quantizer of size $\triangle$ is used, is of the order of $\log_2(1/\triangle)$. The number of bits per coefficient is also of the order of $\log_2(1/\triangle)$, so the rate R is of the order

$$R \sim \log_2^2(1/\triangle) \sim J^2. \qquad (61)$$

The distortion or squared error is proportional to $\triangle^2$ (for each coefficient), times the number of scales or, using $\triangle = 2^{-J}$

$$D \sim J \cdot 2^{-J}. \qquad (62)$$

Using (61), we get

$$D_W(R) = C_1 \cdot \sqrt{R} \cdot 2^{-C_2\sqrt{R}} \qquad (63)$$

for the distortion-rate behavior of a wavelet scheme. Note that we ignored the cost of indexing the location. This cost turns out to be quite small (order J), because the coefficients are all gathered around the discontinuity. For more general signals, Prandoni [31] has shown that non-linear approximation of piecewise polynomials with maximal degree N, using a wavelet compression where the wavelet has $N+1$ zero moments, leads to

$$D_W(R) = C_W'' \left(1 + C_W'\sqrt{R}\right) \cdot 2^{-\sqrt{C_W R}} \qquad (64)$$

where the constants depend on N and the number of discontinuities. The disturbing news is the $\sqrt{R}$ term in the exponent, which is very far from the expected high rate

While a powerful *N*-term approximation is desirable, it must be followed by appropriate compression.

behavior. Such behavior has been shown more generally for piecewise smooth functions by Cohen et al. [5]. (Note that this is a much broader class than piecewise polynomials.) A direct approach to compression of piecewise polynomials, based on an oracle telling us where discontinuities are, leads to

$$D_p(R) = C'_p \cdot 2^{-C_p \cdot R}, \tag{65}$$

and such behavior is achievable using dynamic programming [31]. Clearly (65) is much superior to (64) at high rates. Now, why is the wavelet compression performing suboptimally? The intuitive reason is simple: by compressing the coefficients across scales independently, one

does not take advantage of the interscale dependencies. Several attempts have been made to use such dependencies. For example, zero trees [37] predict lack of energy across scales, and hidden Markov models [8] indicate dependencies across scales. In the case of piecewise polynomials with discontinuities, the behavior across scales is actually deterministic, and this can be modeled using wavelet footprints [16]. Using footprints on a wavelet decomposition of a piecewise polynomial allows one to achieve the best possible behavior given in (65), while using a wavelet transform. Examples of using footprints for denoising can be seen in Fig. 11.

The Two-Dimensional Case

Given the good performance of wavelets for piecewise smooth functions in one dimension, one would hope for good results in two dimensions as well.(Here, we consider separable wavelets in two dimensions, which are commonly used in practice.) Unfortunately, such is not the case. In essence, wavelets are good at catching zero-dimensional singularities, but two-dimensional
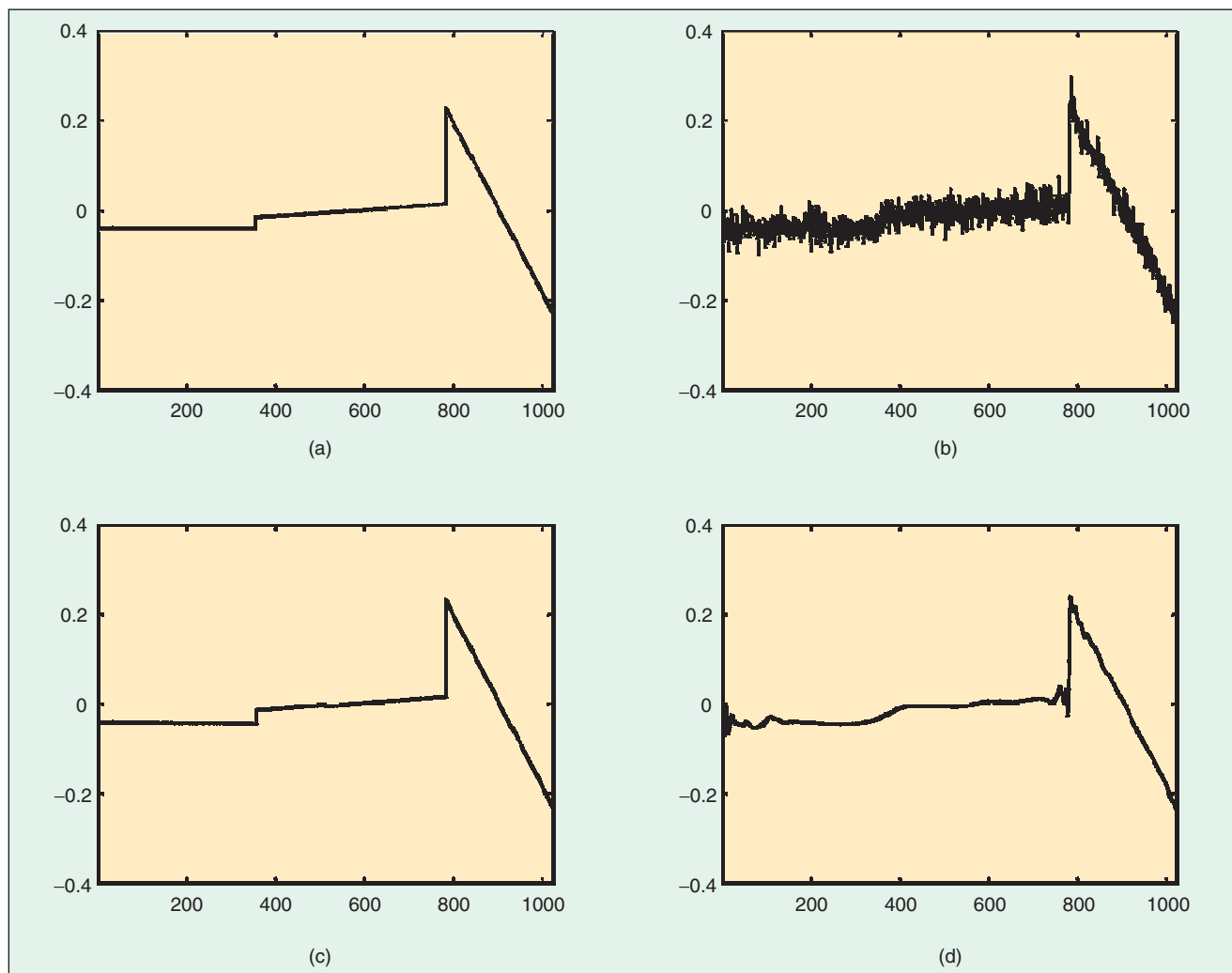
▲ 11. Denoising in wavelet domain using thresholding and footprints. (a) Original. (b) Noisy version (SNR = 11.1 dB). (c) Denoising using footprints (SNR = 31.4 dB). (d) Denoising using standard wavelet thresholding (SNR = 20.1 dB).
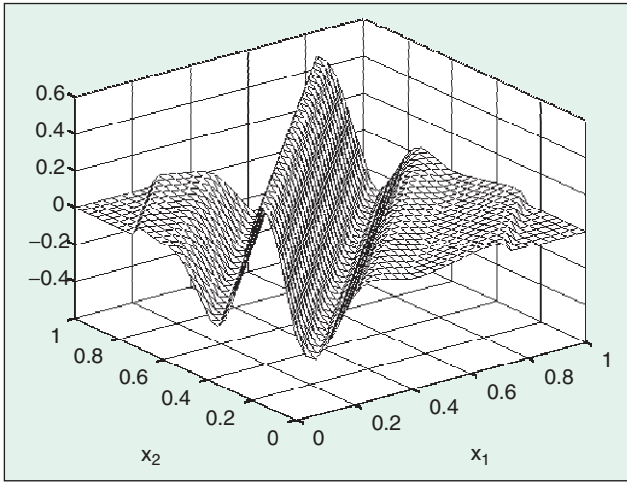
▲ *12. Typical ridgelet, at angle θ with a profile given by $\psi_{a,b}(t)$.*
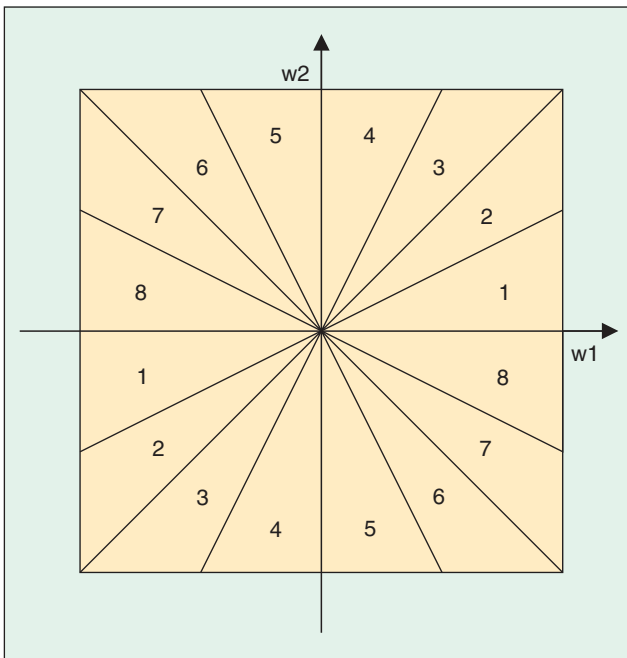
▲ *13. Frequency division of a directional filter bank with 8 bands.*

piecewise smooth signals resembling images have one-dimensional singularities. That is, smooth regions are separated by edges but edges themselves are typically smooth curves. Intuitively, wavelets will be good at isolating the discontinuity orthogonal to the edge, but will not see the smoothness along the edge. This can be verified by looking at the approximation power of a two-dimensional wavelet basis. Such a basis is obtained by a tensor-product of one dimensional wavelets. An analysis similar to that in the "Linear versus Noninear Approximation" Section shows that the M-term nonlinear approximation of a simple piecewise constant function with a linear discontinuity leads to a quadratic error of the order

$$\widetilde{\varepsilon}_M \sim 1/M. \qquad (66)$$

This disappointing behavior indicates that more powerful bases are needed in higher dimensions.

True Two-Dimensional Bases

As the wavelet example shows, separable bases are not suited for "true" two-dimensional objects. What is needed are transforms and bases that include some form of "geometry" and that are truly two dimensional. (The notion of geometry is not easy to formalize in our context, but the intuition is that the dimensions are not independent, and certain shapes are more likely than others.)

Besides the two-dimensional Fourier and wavelet transforms, which are both separable, the Radon transform plays a key role. This transform, studied early in the 20th century [33], was rediscovered several times in fields ranging from astronomy to medical imaging (see [11] for an excellent overview). The Radon transform maps a function $f(x,y)$ into $RA_f(\theta,t)$ by taking line integrals at angle θ and location t

$$RA_f(\theta,t) = \int \int f(x,y)\delta(x\cos\theta + y\sin\theta - t)dxdy. \qquad (67)$$

A key insight to construct directional bases from the Radon transform was provided by Candès and Donoho [2], [3] with the ridgelet transform. The idea is to map a one-dimensional singularity, like a straight edge, into a point singularity using the Radon transform. Then, the wavelet transform can be used to handle the point singularity. To develop the intuition, consider a two-dimensional Heaviside-like function which is −1 on the left of a line $\delta(x\cos\theta_0 + y\sin\theta_0 - t_0)$, and 1 on the right. This is an "infinite" edge function of angle θ_0 and location t_0.

Because of the infinite extent, we cannot take the integral in (67) without additional constraints (like applying a smooth window), but intuitively, for any angle $\theta \neq \theta_0$, the projection $RA_f(\theta,t)$ is smooth in t, while for θ_0, the result is a one-dimensional Heaviside function in t with a singularity in t_0. It is thus natural to take a wavelet transform along t, leading to the definition of the continuous ridgelet transform as [3]

$$RI_f(a,b,\theta) = \int \psi_{a,b}(t)RA_f(\theta,t)dt, \qquad (68)$$

where $\psi_{a,b}(t) = a^{-1/2} \psi(t - b/a)$ and $\psi(t)$ is a wavelet with at least one zero moment. The "atoms" of analysis are infinite ridges at an angle θ, location b and scale a, where the profile of the ridge is given by the wavelet. One such ridgelet is shown in Fig. 12.

The transform in (68) is continuous in all parameters. Appropriate discretization and localization leads to sets of localized directional ridges that can efficiently represent two-dimensional functions with edge-like discontinuities (more precisely, one can show that a frame representation is obtained [3]).

It is also possible to develop discrete schemes that work directly on finite size, sampled data, while emulating the principles of the ridgelet transform. Such schemes typically implement a discrete Radon transform in space or Fourier domain (in the latter case, using the projection slice theorem) followed by a discrete wavelet transform. One exam-

ple is given in [40], where ridgelet frames are constructed for image denoising. Another example is the construction of an orthonormal ridgelet basis from the finite Radon transform and the discrete wavelet transform [13].

The important point with ridgelets is that unlike wavelets, they achieve fast N-term approximation of objects with straight edges [3]. For the example of a Heaviside function, one gets an M-term nonlinear approximation with

$$\widetilde{\varepsilon}_M \sim 2^{-M}. \tag{69}$$

Therefore, they can be used as building blocks in more complex schemes to approximate objects with smooth edges. One such scheme uses localized ridgelets of appropriate size along edges (e.g., with fixed or adaptive blocksizes [40]).

Ridgelets can be combined with multiresolution schemes to get multiscale ridgelets. Combined with bandpass filtering, this gives rise to curvelets [4].
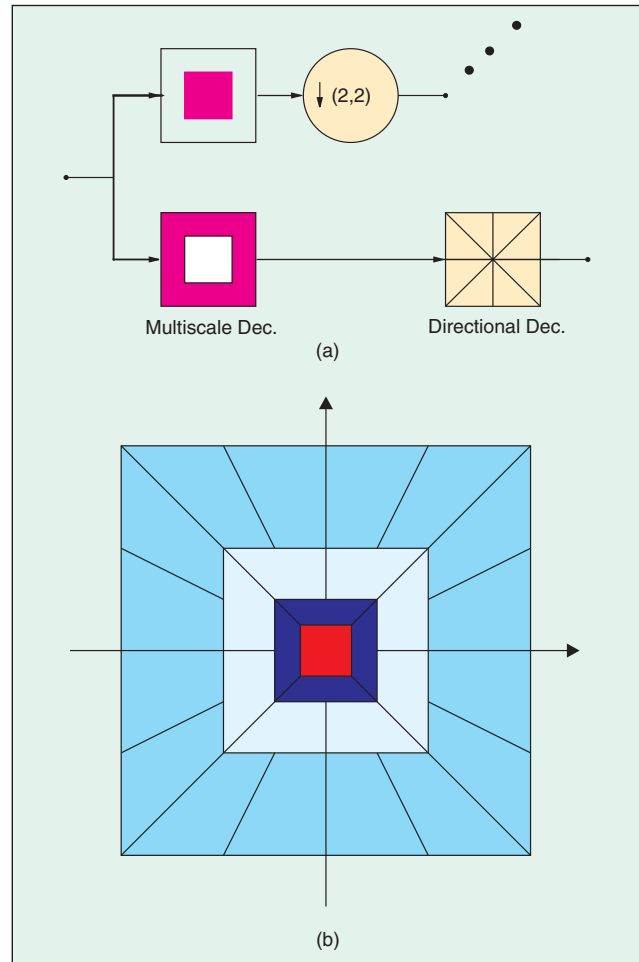
Directional Filter Banks

To get directional analysis, one can alternatively use directional filter banks [1]. In such a case, the basis functions are given by the filter impulse responses and their translates with respect to the subsampling grid. Such filter banks can be designed directly, or through iteration of elementary filter banks. They lead to bases if critically subsampled, or frames if oversampled.

Fig. 13 shows the frequency division achieved by an ideal directional filter bank with eight channels. When this is combined with a multiresolution, pyramidal decomposition, one obtains a curvelet-like decomposition [4], [14]. Specifically, a pyramidal decomposition into bandpass channels (see Fig. 14(a)) is followed by a directional analysis of the bandpass channels. The number of directions is increased as frequency increases, and the resulting frequency split is shown in Fig. 14(b). This system is called a pyramidal directional filter bank (PDFB). A test image and its decomposition is shown in Fig. 15, showing how different directions are separated, and this in a multiresolution manner.

As an example application, a simple denoising algorithm (based on thresholding the small magnitude coefficients in the decomposition) is applied on both a wavelet decomposition (this is a standard denoising procedure) and a pyramidal directional filter bank decomposition. As can be seen in Fig. 16, the PDFB catches directionality more efficiently, producing more pleasing visual results and better SNR performance. This indicates the potential of such nonseparable, directional multiresolution schemes.

Two-Dimensional Bases and Compression

As we had seen in one dimension, a good N-term approximation is not yet a guarantee for good compression. While a powerful N-term approximation is desirable, it must be followed by appropriate compression. Thus, the
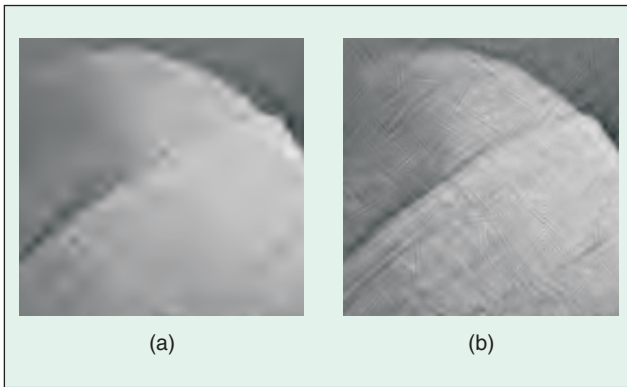

▲ 14. Pyramidal directional filter bank. (a) A standard multiscale decomposition into octave bands, where the lowpass channel is subsampled while the highpass is not. (b) Resulting frequency division, where the number of directions is increased with frequency.
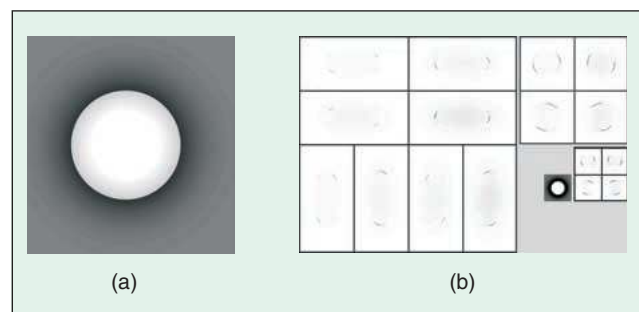
▲ 15. Pyramidal directional filter bank decomposition. (a) Simple test image. (b) Decomposition into directional bandpass images (8, 4 and 4) and a lowpass image.

topic of compression of two-dimensional piecewise smooth functions is still quite open. Several promising approaches are currently under investigation, including compression in ridgelet and curvelet domain, compression along curves using "bandelets" [30] and generalization of footprints in two dimensions or edgeprints [17].

▲ *16. Comparison of threshold-based denoising methods using wavelets (a) and a pyramidal directional filter bank (b). The SNR is 13.82 dB and 15.42 dB, respectively.*

Conclusion

The interplay of representation, approximation, and compression of signals was reviewed. For piecewise smooth signals, we showed the power of wavelet-based methods, in particular for the one-dimensional case. For two-dimensional signals, where wavelets do not provide the answer for piecewise smooth signals with curve singularities, new approaches and open problems were indicated. Such approaches rely on new bases with potentially high impact on image processing, for such problems as denoising, compression and classification.

Acknowledgments

Martin Vetterli received his Engineering degree from Eidgenoessische Technische Hochschule Zuerich (ETHZ), his M.S. from Stanford University, and his Doctorate from Ecole Polytechnique Federale de Lausanne (EPFL). He was with Columbia University, New York, and the University of California at Berkeley before joining the Communication Systems Department of the Swiss Federal Institute of Technology in Lausanne. His research interests include signal processing and communications, in particular, wavelet theory and applications, image and video compression, joint source-channel coding, and self-organized communication systems. He has has won best paper awards from EURASIP in 1984 and the IEEE Signal Processing Society in 1991 and 1996. He is the co-author, with J. Kovačević, of the textbook *Wavelets and Subband Coding* (Prentice-Hall, 1995).

References

[1] R.H. Bamberger and M.J.T. Smith, "A filter bank for the directional decomposition of images: Theory and design," *IEEE Trans. Signal Processing*, vol. 40., pp. 882-893, Apr. 1992.

[2] E. Candès, "Ridgelets: Theory and applications," Ph.D. dissertation, Dept. Statistics, Stanford University, Stanford, CA, 1998.

[3] E. Candès and D.L. Donoho, "Ridgelets: A key to higher-dimensional intermittency?," *Phil. Trans. R. Soc. London A*., pp. 2495-2509, 1999.

[4] E.J. Candès and D.L. Donoho, "Curvelets—A surprisingly effective nonadaptive representation for objects with edges," in *Curve and Surface Fitting*, A. Cohen, C. Rabutand, and L.L. Schumaker, Eds. Saint-Malo: Vanderbilt University Press, 1999.

[5] A. Cohen, W. Dahmen, and I. Daubechies, "Tree approximation and optimal encoding," *Appl. Computational Harmonic Anal.,* to be published.

[6] R.R. Coifman and M.V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Inform. Theory (Special Issue on Wavelet Transforms and Multiresolution Signal Analysis)*, vol. 38, pp. 713-718, Mar. 1992.

[7] T.M. Cover and J.A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[8] M. Crouse, R.D. Nowak, and R.G. Baraniuk, "Wavelet-based signal processing using hidden Markov models," *IEEE Trans. Signal Processing (Special Issue on Wavelets and Filterbanks)*, pp. 886-902, Apr. 1998.

[9] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Commun. Pure Appl. Math.*, vol. 41, pp. 909-996, Nov. 1988.

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

[11] S.R. Deans, *The Radon Transform and Some of its Applications*. New York: Wiley, 1983.

[12] R.A. DeVore, B. Jawerth, and B. J. Lucier, "Image compression through wavelet transform coding," *IEEE Trans. Inform. Theory (Special Issue on Wavelet Transforms and Multiresolution Signal Analysis)*, vol. 38, pp. 719-746, Mar. 1992.

[13] M. Do and M. Vetterli, "Orthonormal finite ridgelet transform for image compression," in *Proc. IEEE Int. Conf. Image Processing, ICIP 2000,* Vancouver, Canada, Sept. 2000, pp. 367-370.

[14] M. Do and M. Vetterli, "Pyramidal directional filter banks and curvelets," in *Proc. IEEE Int. Conf. Image Processing, ICIP 2001,* Patras, Greece, Oct. 2001.

[15] D. Donoho, M. Vetterli, R. DeVore, and I Daubechies, "Data compression and harmonic analysis," *IEEE Trans. Inform Theory (Special Issue, Information Theory: 1948-1998 Commemorative Issue)*, vol. 44, pp. 2435-2476, Oct. 1998.

[16] P.L. Dragotti and M. Vetterli, "Wavelet transform footprints: Catching singularities for compression and denoising," in *Proc. IEEE Int. Conf. Image Processing, ICIP 2000,* Vancouver, Canada, Sept. 2000, pp. 363-366.

[17] P.L. Dragotti and M. Vetterli, "Footprints and edgeprints for image denoising and compression," in *Proc. IEEE Int. Conf. Image Processing, ICIP 2001,* Patras, Greece, Oct. 2001.

[18] J. Fourier, *Théorie Analytique de la Chaleur*. Paris, France: Gauthier-Villars, 1888.

[19] A. Gersho and R.M. Gray, *Vector Quantization and Signal Compression*. Norwell, MA: Kluwer, 1992.

[20] V.K. Goyal, "Theoretical foundations of transform coding," *IEEE Signal Processing Mag.*, vol. 18, pp. 9-21, Sept. 2001

[21] V.K. Goyal, *Transform Coding*. SIAM, to be published.

[22] V.K. Goyal, J. Zhuang, and M. Vetterli, "Transform coding with backward adaptive updates," *IEEE Trans. Inform. Theory*, to be published.

[23] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Recognition Machine Intell.*, vol. 11, pp. 674-693, July 1989.

[24] S. Mallat, *A Wavelet Tour of Signal Processing*. San Diego, CA: Academic, 1998.

[25] S. Mallat and W.L. Hwang, "Singularity detection and processing with wavelets," *IEEE Trans. Inform. Theory (Special Issue on Wavelet Transforms and Multiresolution Signal Analysis)*, vol. 38, pp. 617-643, Mar. 1992.

[26] S.G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing (Special Issue on Wavelets and Signal Processing)*, vol. 41, pp. 3397-3415, Dec. 1993.

[27] F. Mintzer, "Filters for distortion-free two-band multirate filter banks," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 33, pp. 626-630, June 1985.

[28] A.V. Oppenheim and R.W. Schafer, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[29] A. Ortega, K. Ramchandran, and M. Vetterli, "Optimal trellis-based buffered compression and fast approximations," *IEEE Trans. Image Processing*, vol. 3, pp. 26-40, Jan. 1994.

[30] E. Le Pennec and S. Mallat, "Image compression with geometric wavelets," in *Proc. IEEE Int. Conf. Image Processing, ICIP 2000,* Vancouver, Canada, Sept. 2000, pp. 661-664.

[31] P. Prandoni, "Optimal segmentation techniques for piecewise stationary signals," Ph.D. dissertation, EPFL, Communications Systems, June 1999.

[32] P. Prandoni and M. Vetterli, "Approximation and compression of piecewise-smooth functions," *Phil. Trans. Roy. Soc. London*, vol. 357, p. 1760, Sept. 1999.

[33] J. Radon, "Ueber die bestimmung von funktionen durch ihre integralwerte längst gewisser mannigfaltigkeiten," *Berichte Sächsische Akademie der Wissenschaften, Leipzig*, pp. KI. 69, 262-267, 1917.

[34] K. Ramchandran and M. Vetterli, "Best wavelet packet bases in a rate-distortion sense," *IEEE Trans. Image Processing*, vol. 2, pp. 160-175, Apr. 1993.

[35] A. Said and W.A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits Syst. Video*, vol. 6, pp. 243-249, June 1996.

[36] C.B. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, 1948.

[37] J.M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Processing (Special Issue, Wavelets and Signal Processing)*, vol. 41, pp. 3445-3462, Dec. 1993.

[38] T. Skodras, C. Christopoulos, and T. Ebrahimi, "The JPEG 2000 still image compression standard," *IEEE Signal Processing Mag.*, vol. 18, pp. 36-58, Sept. 2001.

[39] M.J.T. Smith and T.P. Barnwell III, "Exact reconstruction for tree-structured subband coders," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 34, pp. 431-441, June 1986.

[40] J.L. Starck, E. Candès, and D. Donoho, "The curvelet transform for image denoising," *IEEE Trans. Image Processing*, submitted for publication.

[41] G. Strang and T. Nguyen, *Wavelets and Filter Banks*. Cambridge, MA: Wellesley-Cambridge, 1996.

[42] B. Usevitch, "Wavelet-based image compression," *IEEE Signal Processing Mag.*, vol. 18, pp. 22-35, Sept. 2001.

[43] P.P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[44] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[45] Claudio Weidmann, "Oligoquantization in low-rate lossy source coding," Ph.D. dissertation, EPFL, Communication Systems, July 2000.