Data Compression and Harmonic Analysis

David L. Donoho, Martin Vetterli, *Fellow, IEEE*, R. A. DeVore, and Ingrid Daubechies, *Senior Member, IEEE*

(Invited Paper)

***Abstract*— In this paper we review some recent interactions between harmonic analysis and data compression. The story goes back of course to Shannon's $R(D)$ theory in the case of Gaussian stationary processes, which says that transforming into a Fourier basis followed by block coding gives an optimal lossy compression technique; practical developments like transform-based image compression have been inspired by this result. In this paper we also discuss connections perhaps less familiar to the Information Theory community, growing out of the field of harmonic analysis. Recent harmonic analysis constructions, such as wavelet transforms and Gabor transforms, are essentially optimal transforms for transform coding in certain settings. Some of these transforms are under consideration for future compression standards.**

We discuss some of the lessons of harmonic analysis in this century. Typically, the problems and achievements of this field have involved goals that were not obviously related to practical data compression, and have used a language not immediately accessible to outsiders. Nevertheless, through an extensive generalization of what Shannon called the "sampling theorem," harmonic analysis has succeeded in developing new forms of functional representation which turn out to have significant data compression interpretations. We explain why harmonic analysis has interacted with data compression, and we describe some interesting recent ideas in the field that may affect data compression in the future.

***Index Terms*— Besov spaces, block coding, cosine packets, ϵ-entropy, Fourier transform, Gabor transform, Gaussian process, Karhunen–Loève transform, Littlewood–Paley theory, non-Gaussian process, n-widths, rate-distortion, sampling theorem, scalar quantization, second-order statistics, Sobolev spaces, subband coding, transform coding, wavelet packets, wavelet transform, Wilson bases.**

"Like the vague sighings of a wind at even
That wakes the wavelets of the slumbering sea."
Shelley, 1813

Manuscript received June 1, 1998; revised July 6, 1998. The work of D. L. Donoho was supported in part by NSF under Grant DMS-95-05151, by AFOSR under Grant MURI-95-F49620-96-1-0028, and by other sponsors. The work of M. Vetterli was supported in part by NSF under Grant MIP-93-213002 and by the Swiss NSF under Grant 20-52347.97. The work of R. A. DeVore was supported by ONR under Contract N0014-91-J1343 and by Army Research Office under Contract N00014-97-0806. The work of I. Daubechies was supported in part by NSF under Grant DMS-9706753, by AFOSR under Grant F49620-98-1-0044, and by ONR under Contract N00014-96-1-0367.

D. L. Donoho is with Stanford University, Stanford, CA USA 94205.

M. Vetterli is with the Communication Systems Division, the Swiss Federal Institute of Technology, CH-1015 Lausanne, Switzerland, and with the Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720 USA.

R. A. DeVore is with the Department of Mathematics, University of South Carolina, Columbia, SC 29208 USA.

I. Daubechies is with the Department of Mathematics, Princeton University, Princeton, NJ 08544 USA.

Publisher Item Identifier S 0018-9448(98)07005-9.

"ASBOKQTJEL"

Postcard from J. E. Littlewood to A. S. Besicovich, announcing A. S. B.'s election to fellowship at Trinity

"... the 20 bits per second which, the psychologists assure us, the human eye is capable of taking in, ..."

D. Gabor, Guest Editorial, *IRE Trans. Inform Theory*, Sept. 1959.

I. Introduction

DATA compression is an ancient activity; abbreviation and other devices for shortening the length of transmitted messages have no doubt been used in every human society. Language itself is organized to minimize message length, short words being more frequently used than long ones, according to Zipf's empirical distribution.

Before Shannon, however, the activity of data compression was informal and *ad hoc*; Shannon created a formal intellectual discipline for both lossless and lossy compression, whose 50th anniversary we now celebrate.

A remarkable outcome of Shannon's formalization of problems of data compression has been the intrusion of sophisticated theoretical ideas into widespread use. The JPEG standard, set in the 1980's and now in use for transmitting and storing images worldwide, makes use of quantization, run-length coding, entropy coding, and fast cosine transformation. In the meantime, software and hardware capabilities have developed rapidly, so that standards currently in process of development are even more ambitious. The proposed standard for still image compression—JPEG-2000—contains the possibility for conforming codecs to use trellis-coded quantizers, arithmetic coders, and fast wavelet transforms.

For the authors of this paper, one of the very striking features of recent developments in data compression has been the applicability of ideas taken from the field of harmonic analysis to both the theory and practice of data compression. Examples of this applicability include the appearance of the fast cosine transform in the JPEG standard and the consideration of the fast wavelet transform for the JPEG-2000 standard. These fast transforms were originally developed in applied mathematics for reasons completely unrelated to the demands of data compression; only later were applications in compression proposed.

John Tukey became interested in the possibility of accelerating Fourier transforms in the early 1960's in order to enable spectral analysis of long time series; in spite of the fact that Tukey coined the word "bit," there was no idea in his mind at the time of applications to data compression. Similarly,

the construction of smooth wavelets of compact support was prompted by questions posed implicitly or explicitly by the multiresolution analysis concept of Mallat and Meyer, and not, at that time, by direct applications to compression.

In asking about this phenomenon—the applicability of computational harmonic analysis to data compression—there are, broadly speaking, two extreme positions to take.

The first, *maximalist* position holds that there is a deep reason for the interaction of these disciplines, which can be explained by appeal to information theory itself. This point of view holds that sinusoids and wavelets will necessarily be of interest in data compression because they have a special "optimal" role in the representation of certain stochastic processes.

The second, *minimalist* position holds that, in fact, computational harmonic analysis has exerted an influence on data compression merely by happenstance. This point of view holds that there is no fundamental connection between, say, wavelets and sinusoids, and the structure of digitally acquired data to be compressed. Instead, such schemes of representation are privileged to have fast transforms, and to be well known, to have been well studied and widely implemented at the moment that standards were being framed.

When one considers possible directions that data compression might take over the next fifty years, the two points of view lead to very different predictions. The maximalist position would predict that there will be continuing interactions between ideas in harmonic analysis and data compression; that as new representations are developed in computational harmonic analysis, these will typically have applications to data compression practice. The minimalist position would predict that there will probably be little interaction between the two areas in the future, or that what interaction does take place will be sporadic and opportunistic.

In this paper, we would like to give the reader the background to appreciate the issues involved in evaluating the two positions, and to enable the reader to form his/her own evaluation. We will review some of the connections that have existed classically between methods of harmonic analysis and data compression, we will describe the disciplines of theoretical and computational harmonic analysis, and we will describe some of the questions that drive those fields.

We think there is a "Grand Challenge" facing the disciplines of both theoretical and practical data compression in the future: the challenge of dealing with the particularity of naturally occurring phenomena. This challenge has three facets:

GC1 Obtaining accurate models of naturally occurring sources of data.

GC2 Obtaining "optimal representations" of such models.

GC3 Rapidly computing such "optimal representations."

We argue below that current compression methods might be far away from the ultimate limits imposed by the underlying structure of specific data sources, such as images or acoustic phenomena, and that efforts to do better than what is done today—particularly in specific applications areas—are likely to pay off.

Moreover, parsimonious representation of data is a fundamental problem with implications reaching well beyond compression. Understanding the compression problem for a given data type means an intimate knowledge of the modeling and approximation of that data type. This in turn can be useful for many other important tasks, including classification, denoising, interpolation, and segmentation.

The discipline of harmonic analysis can provide interesting insights in connection with the Grand Challenge.

The history of theoretical harmonic analysis in this century repeatedly provides evidence that in attacking certain challenging and important problems involving characterization of infinite-dimensional classes of functions, one can make progress by developing new functional representations based on certain geometric analogies, and by validating that those analogies hold in a quantitative sense, through a norm equivalence result. Also, the history of computational harmonic analysis has repeatedly been that geometrically motivated analogies constructed in theoretical harmonic analysis have often led to fast concrete computational algorithms.

The successes of theoretical harmonic analysis are interesting from a data compression perspective. What the harmonic analysts have been doing—showing that certain orthobases afford certain norm equivalences—is analogous to the classical activity of showing that a certain orthobasis diagonalizes a quadratic form. Of course, the diagonalization of quadratic forms is of lasting significance for data compression in connection with transform coding of Gaussian processes. So one could expect the new concept to be interesting *a priori*. In fact, the new concept of "diagonalization" obtained by harmonic analysts really does correspond to transform coders—for example, wavelet coders and Gabor coders.

The question of whether the next 50 years will display interactions between data compression and harmonic analysis more like a maximalist or a minimalist profile is, of course, anyone's guess. This paper provides encouragement to those taking the maximalist position.

The paper is organized as follows. At first, classical results from rate-distortion theory of Gaussian processes are reviewed and interpreted (Sections II and III). In Section IV, we develop the functional point of view, which is the setting for harmonic analysis results relevant to compression, but which is somewhat at variance with the digital signal processing viewpoint. In Section V, the important concept of Kolmogorov ϵ-entropy of function classes is reviewed, as an alternate approach to a theory of compression. In Section VI, practical transform coding as used in image compression standards is described. We are now in a position to show commonalities between the approaches seen so far (Section VII), and then to discuss limitations of classical models (Section VIII) and propose some variants by way of simple examples. This leads to pose the "Grand Challenges" to data compression as seen from our perspective (Section IX), and to overview how Harmonic Analysis might participate in their solutions. This leads to a survey of Harmonic Analysis results, in particular on norm equivalences (Sections XI–XIII) and nonlinear approximation (Section XIV). In effect, one can show that harmonic analysis, which is effective at establishing norm

equivalences, leads to coders which achieve the ϵ-entropy of functional classes (Section XV). This has a transform coding interpretation (Section XVI), showing a broad analogy between the deterministic concept of unconditional basis and the stochastic concept of Karhunen–Loève expansion. In Section XVII, we discuss the role of tree-based ideas in harmonic analysis, and the relevance for data compression. Section XVIII briefly surveys some harmonic analysis results on time–frequency-based methods of data compression. The fact that many recent results from theoretical harmonic analysis have computationally effective counterparts is described in Section XIX. Practical coding schemes using or having led to some of the ideas described thus far are described in Section XX, including current advanced image compression algorithms based on fast wavelet transforms. As a conclusion, a few key contributors in harmonic analysis are used to iconify certain key themes of this paper.

II. $R(D)$ for Gaussian Processes

Fifty years ago, Claude Shannon launched the subject of lossy data compression of continuous-valued stochastic processes [83]. He proposed a general idea, the rate-distortion function, which in concrete cases (Gaussian processes) leads to a beautifully intuitive method to determine the number of bits required to approximately represent sample paths of a stochastic process.

Here is a simple outgrowth of the Shannon theory, important for comparison with what follows. Suppose $X(t)$ is a Gaussian zero-mean stochastic process on an interval T and let $N(D,X)$ denote the minimal number of codewords needed in a codebook $\mathcal{C} = \{X'\}$ so that

$$E \min_{X' \in \mathcal{C}} \|X - X'\|^2_{L^2(T)} \le D. \tag{2.1}$$

Then Shannon proposed that in an asymptotic sense

$$\log N(D,X) \approx R(D,X) \tag{2.2}$$

where $R(D,X)$ is the rate-distortion function for X

$$R(D,X) = \inf \{I(X,Y):\ E\|X-Y\|^2_{L^2(T)} \le D\} \tag{2.3}$$

with $I(X,Y)$ the usual mutual information, given formally by

$$I(X,Y) = \int p(x,y) \log \frac{p(x,y)}{p(x)p(y)}\, dx\, dy. \tag{2.4}$$

Here $R(D,X)$ can be obtained in parametric form from a formula which involves functionals of the covariance kernel $K(s,t) = \mathrm{Cov}\,(X(t),X(s))$; more specifically of the eigenvalues (λ_k). In the form first published by Kolmogorov (1956), but probably known earlier to Shannon, for $\theta > 0$ we have

$$R(D_\theta) = \sum_{k,\lambda_k > \theta} \log\,(\lambda_k/\theta) \tag{2.5}$$

where

$$D_\theta = \sum_k \min\,(\theta, \lambda_k). \tag{2.6}$$

The random process Y^* achieving the minimum of the mutual information problem can be described as follows. It has a covariance with the same eigenfunctions as that of X, but the eigenvalues are reduced in size:

$$\mu_k = (\lambda_k - \theta)_+.$$

To obtain a codebook achieving the value predicted by $R(D,X)$, Shannon's suggestion would be to sample realizations from the reproducing process Y^* realizing the minimum of the least mutual information problem.

Formally, then, the structure of the optimal data compression problem is understood by passing to the Karhunen–Loève expansion

$$X(t) = \sum_k \sqrt{\lambda_k} Z_k \phi_k(t).$$

In this expansion, the coefficients are independent zero-mean Gaussian random variables. We have a similar expansion for the reproducing distribution

$$Y^*(t) = \sum_k \sqrt{\mu_k} \tilde{Z}_k \phi_k(t).$$

The process Y^* has only finitely many nonzero coefficients, namely, those coefficients at indices k where $\lambda_k > \theta$; let $K(D)$ denote the indicated subset of coefficients. Random codebook compression is effected by comparing the vector of coefficients $(\langle X, \phi_k \rangle:\ k \in K(D))$ with a sequence of codewords $(\sqrt{\mu_k}\tilde{Z}_{k,i}:\ k \in K(D))$, for $i = 1, \cdots, N$, looking for a closest match in Euclidean distance. The approach just outlined is often called "reverse waterfilling," since only coefficients above a certain water level are described in the coding process.

As an example, let $T = [0,1)$ and let $X(t)$ be the Brownian Bridge, i.e., the continuous Gaussian process with covariance $K(s,t) = \min(s,t) - st$. This Gaussian process has $X(0) = X(1) = 0$ and can be obtained by taking a Brownian motion $B(t)$ and "pinning" it down at 1: $X(t) = B(t) - tB(1)$. The covariance kernel has sinusoids for eigenvectors: $\phi_k(t) = \sin(2\pi kt)$, and has eigenvalues $\lambda_k = (4\pi^2 k^2)^{-1}$. The subset $K(D)$ amounts to a frequency band of the first $\#K(D) \asymp D^{-1}$ frequencies as $D \to 0$. (Here and below, we use $A \asymp B$ to mean that the two expressions are equivalent to within multiplicative constants, at least as the underlying argument tends to its limit.) Hence the compression scheme amounts to "going into the frequency domain," "extracting the low frequency coefficients," and "comparing with codebook entries." The number of low frequency coefficients to keep is directly related to the desired distortion level. The achieved $R(D,X)$ in this case scales as $R(D,X) \asymp D^{-1}$.

Another important family of examples is given by stationary processes. In this case, the eigenfunctions of the covariance are essentially sinusoids, and the Karhunen–Loève expansion has a more concrete interpretation. To make things simple and analytically exact, suppose we are dealing with the circle $T = [0, 2\pi)$, and considering stationarity with respect to circular shifts. The stationarity condition is $K(s,t) = \gamma(s -_\circ t)$, where $s -_\circ t$ denotes circular (clock) arithmetic. The eigenfunctions

of K are the sinusoids

$$\phi_1(t) = 1/\sqrt{2\pi}$$
$$\phi_{2k}(t) = \cos(kt)/\sqrt{\pi}$$
$$\phi_{2k+1}(t) = \sin(kt)/\sqrt{\pi}$$

and the eigenvalues λ_k are the Fourier coefficients of γ

$$\lambda_k = \int_0^{2\pi} \gamma(t)\phi_k(t)\,dt.$$

We can now identify the index k with frequency, and the Karhunen–Loève representation effectively says that the Fourier coefficients of X are independently zero-mean Gaussian, with variances λ_k, and that the reproducing process Y^* has Fourier coefficients which are independent Gaussian coefficients with variances μ_k. For instance, consider the case $\lambda_k \sim Ck^{-2m}$ as $k \to \infty$; then the stationary process has nearly $(m - 1/2)$-derivatives in a mean-square sense. For this example, the band $K(D)$ amounts to the first $\#K(D) \asymp D^{-1/(2m-1)}$ frequencies. Hence once again the compression scheme amounts to "going into the frequency domain," "extracting the low frequency coefficients," and "comparing with codebook entries," and the number of low frequency coefficients retained is given by the desired distortion level. The achieved $R(D, X)$ in this case scales as $R(D, X) \asymp D^{-1/(2m-1)}$.

III. Interpretations of $R(D)$

The compression scheme described by the solution of $R(D)$ in the Gaussian case has several distinguishing features.

- *Transform Coding Interpretation:* Undoubtedly the most important thing to read off from the solution is that *data compression can be factored into two components: a transform step followed by a coding step*. The transform step takes continuous data and yields discrete sequence data; the coding step takes an initial segment of this sequence and compares it with a codebook, storing the binary representation of the best matching codeword.
- *Independence of the Transformed Data:* The transform step yields uncorrelated Gaussian data, hence *stochastically independent* data, which are, after normalization by factors $1/\sqrt{\lambda_k}$, identically distributed. Hence, the apparently abstract problem of coding a process $X(t)$ becomes closely related to the concrete problem of coding a Gaussian memoryless source, under weighted distortion measure.
- *Manageable Structure of the Transform:* The transform itself is mathematically well-structured. It amounts to expanding the object X in the orthogonal eigenbasis associated with a self-adjoint operator, which at the abstract level is a well-understood notion. In certain important concrete cases, such as the examples we have given above, the basis even reduces to a well-known Fourier basis, and so optimal coding involves explicitly *harmonic analysis* of the data to be encoded.

There are two universality aspects we also find remarkable:

- *Universality Across Distortion Level:* The structure of the ideal compression system does not depend on the distortion level; the same transform and coder structure are employed, but details of the "useful components" $K(D)$ change.
- *Universality Across Stationary Processes:* Large classes of Gaussian processes will share approximately the same coder structure, since there are large classes of covariance kernels with the same eigenstructure. For example, all stationary covariances on the circle have the same eigenfunctions, and so, there is a single "universal" transform that is appropriate for transform coding of all such processes—the Fourier transform.

At a higher level of abstraction, we remark on two further aspects of the solution.

- *Dependence on Statistical Characterization:* To the extent that the orthogonal transform is not universal, it nevertheless depends on the statistical characterization of the process in an easily understandable way—via the eigenstructure of the covariance kernel of the process.
- *Optimality of the Basis:* Since the orthobasis underlying the transform step is the Karhunen–Loève expansion, it has an optimality interpretation independently from its coding interpretation; in an appropriate ordering of the basis elements, partial reconstruction from the first K components gives the best mean-squared approximation to the process available from any orthobasis.

These features of the $R(D)$ solution are so striking and so memorable, that it is unavoidable to incorporate these interpretations as deep "lessons" imparted by the $R(D)$ calculation. These "lessons," reinforced by examples such as those we describe later, can harden into a "world view," creating expectations affecting data compression work in practical coding.

- *Factorization:* One expects to approach coding problems by compartmentalization: attempting to design a two-step system, with the first step a transform, and the second step a well-understood coder.
- *Optimal Representation:* One expects that the transform associated with an optimal coder will be an expansion in a kind of "optimal basis."
- *Empiricism:* One expects that this basis is associated with the statistical properties of the process and so, in a concrete application, one could approach coding problems empirically. The idea would be to obtain empirical instances of process data, and to accurately model the covariance kernel (dependence structure) of those instances; then one would obtain the corresponding eigenfunctions and eigenvalues, and design an empirically derived near-ideal coding system.

These expectations are inspiring and motivating. Unfortunately, there is really nothing in the Shannon theory which supports the idea that such "naive" expectations will apply

outside the narrow setting in which the expectations were formed. If the process data to be compressed are not Gaussian, the $R(D)$ derivation mentioned above does not apply, and one has no right to expect these interpretations to apply.

In fact, depending on one's attraction to pessimism, it would also be possible to entertain completely opposite expectations when venturing outside the Gaussian domain. If we consider the problem of data compression of arbitrary stochastic processes, the following expectations are essentially all that one can apply.

- *Lack of Factorization:* One does not expect to find an ideal coding system for an arbitrary non-Gaussian process that involves transform coding, i.e., a two-part factorization into a transform step followed by a step of coding an independent sequence.
- *Lack of Useful Structure:* In fact, one does not expect to find any intelligible structure whatsoever, in an ideal coding system for an arbitrary non-Gaussian process—beyond the minimal structure on the random codebook imposed by the $R(D)$ problem.

For purely human reasons, it is doubtful, however, that this set of "pessimistic" expectations is very useful as a working hypothesis. The more "naive" picture, taking the $R(D)$ story for Gaussian processes as paradigmatic, leads to the following possibility: *as we consider data compression in a variety of settings outside the strict confines of the original Gaussian $R(D)$ setting, we will discover that many of the expectations formed in that setting still apply and are useful, though perhaps in a form modified to take account of the broader setting*. Thus for example, we might find that factorization of data compression into two steps, one of them an orthogonal transform into a kind of optimal basis, is a property of near-ideal coders that we see outside the Gaussian case; although we might also find that the notion of optimality of the basis and the specific details of the types of bases found would have to change. We might even find that we have to replace "expansion in an optimal basis" by "expansion in an optimal decomposition," moving to a system more general than a basis.

We will see several instances below where the lessons of Gaussian $R(D)$ agree with ideal coders under this type of extended interpretation.

IV. Functional Viewpoint

In this paper we have adopted a point of view we call the *functional viewpoint*. Rather than thinking of data to be compressed as numerical arrays x_u with integer index u, we think of the objects of interest as functions—functions $f(t)$ of time or functions of space $f(x,y)$. To use terminology that Shannon would have found natural, we are considering compression of analog signals. This point of view is clearly the one in Shannon's 1948 introduction of the optimization problem underlying $R(D)$ theory, but it is less frequently seen today, since many practical compression systems start with sampled data. Practiced IT researchers will find one aspect of our discussion unnatural: we study the case where the index set T stays fixed. This seems at variance even with Shannon, who often let the domain of observation grow without bound. The fixed-domain functional viewpoint is essential for developing the themes and theoretical connections we seek to expose in this paper—it is only through this viewpoint that the connections with modern Harmonic analysis become clear. Hence, we pause to explain how this viewpoint can be related to standard information theory and to practical data compression.

A practical motivation for this viewpoint can easily be proposed. In effect, when we are compressing acoustic or image phenomena, there is truly an underlying analog representation of the object, and a digitally sampled object is an approximation to it. Consider the question of an appropriate model for data compression of still-photo images. Over time, consumer camera technology will develop so that standard cameras will routinely be able to take photos with several megapixels per image. By and large, consumers using such cameras will not be changing their photographic compositions in response to the increasing quality of their equipment; they will not be expanding their field of view in picture taking, but rather, they will instead keep the field of view constant, and so as cameras improve they will get finer and finer resolution on the same photos they would have taken anyway. So what is increasing asymptotically in this setting is the resolution of the object rather than the field of view. In such a setting, the functional point of view is sensible. There is a continuum image, and our digital data will sooner or later represent a very good approximation to such a continuum observation. Ultimately, cameras will reach a point where the question of how to compress such digital data will best be answered by knowing about the properties of an ideal system derived for continuum data.

The real reason for growing-domain assumptions in information theory is a technical one: it allows in many cases for the proof of source coding theorems, establishing the asymptotic equivalence between the "formal bit rate" $R(D,X)$ and the "rigorous bit rate" $N(D,X)$. In our setting, this connection is obtained by considering asymptotics of both quantities as $D \to 0$. In fact, it is the $D \to 0$ setting that we focus on here, and it is under this assumption that we can show the usefulness of harmonic analysis techniques to data compression.[1] This may seem at first again at variance with Shannon, who considered the distortion fixed (on a per-unit basis) and let the domain of observation grow without bound.

We have two nontechnical responses.

- *The Future:* With the consumer camera example in mind, high-quality compression of very large data sets may soon be of interest. So the functional viewpoint, and low-distortion coding of the data, may be very interesting settings in the near future.
- *Scaling:* In important situations there is a near equivalence between the "growing domain" viewpoint and the "functional viewpoint." We are thinking here of phenomena like natural images which have scaling properties: if we dilate an image, "stretching it out" to live on a growing

[1] The $D \to 0$ case is usually called the fine quantization or high-resolution limit in quantization theory; see [48].

domain, then after appropriate rescaling, we get statistical properties that are the same as the original image [37], [81]. The relevance to coding is evident, for example, in the stationary Gaussian $R(D)$ case for the process defined in Section II, which has eigenvalues obeying an asymptotic power law, and hence which asymptotically obeys scale invariance at fine scales. Associated to a given distortion level is a characteristic "cutoff frequency" $\#K(D)$; dually this defines a scale of approximation; to achieve that distortion level it is necessary only to know the Fourier coefficients out to frequency $\#K(D)$, or to know the samples of a bandlimited version of the object out to scale $\approx 2\pi/\#K(D)$. This characteristic scale defines a kind of effective pixel size. As the distortion level decreases, this scale decreases, and one has many more "effective pixels." Equivalently, one could rescale the object as a function of D so that the characteristic scale stays fixed, and then the effective domain of observation would grow.

In addition to these relatively general responses, we have a precise response: $D \to 0$ *allows Source Coding Theorems*. To see this, return to the $R(D)$ setting of Section II, and the stochastic process with asymptotic power law eigenvalues given there. We propose grouping frequencies into subbands $K_b = \{k_b, k_b + 1, \cdots, k_{b+1} - 1\}$. The subband boundaries k_b should be chosen in such a way that they get increasingly long with increasing k but that in a relative sense, measured with respect to distance from the origin, they get increasingly narrow

$$\begin{aligned} k_{b+1} - k_b &\to \infty, \qquad b \to \infty \\ (k_{b+1} - k_b)/k_b &\to 1, \qquad b \to \infty. \end{aligned} \tag{4.1}$$

The Gaussian $R(D)$ problem of Section II has the structure suggesting that one needs to code the first $K(D)$ coefficients in order to get a near-ideal system. Suppose we do this by dividing the first $K(D)$ Fourier coefficients of X into subband blocks and then code the subband blocks using the appropriate coder for a block from a Gaussian independent and identically distributed (i.i.d.) source.

This makes sense. For the process we are studying, the eigenvalues λ_k decay smoothly according to a power law. The subbands are chosen so that the variances λ_k are roughly constant in subbands

$$\max\{\lambda_k \colon k \in K_b\} / \min\{\lambda_k \colon k \in K_b\} \to 1, \qquad b \to \infty. \tag{4.2}$$

Within subband blocks, we may then reasonably regard the coefficients as independent Gaussian random variables with a common variance. It is this property that would suggest to encode the coefficients in a subband using a coder for an i.i.d. Gaussian source. The problem of coding Gaussian i.i.d. data is among the most well-studied problems in information theory, and so this subband partitioning reduces the abstract problem of coding the process to a very familiar one.

As the distortion D tends to zero, the frequency cutoff $K(D)$ in the underlying $R(D)$ problem tends to infinity, and so the subband blocks we must code include longer and longer blocks farther and farther out in the frequency domain. These blocks behave more and more nearly like long blocks of Gaussian i.i.d. samples, and can be coded more and more precisely at the rate for a Gaussian source, for example using a random codebook. An increasing fraction of all the bits allocated comes from the long blocks, where the coding is increasingly close to the rate. Hence we get the asymptotic equality of "formal bits" and "rigorous bits" as $D \to 0$.

(Of course in a practical setting, as we will discuss farther below, block coding of i.i.d. Gaussian data, is impractical to "instrument;" there is no known computationally efficient way to code a block of i.i.d. Gaussians approaching the $R(D)$ limit. But in a practical case one can use known suboptimal coders for the i.i.d. problem to code the subband blocks. Note also that such suboptimal coders can perform rather well, especially in the high-rate case, since an entropy-coded uniform scalar quantizer performs within 0.255 bit/sample of the optimum.)

With the above discussion, we hope to have convinced the reader that our functional point of view, although unconventional, will shed some interesting light on at least the high-rate, low-distortion case.

V. The ϵ-Entropy Setting

In the mid-1950's, A. N. Kolmogorov, who had been recently exposed to Shannon's work, introduced the notion of the ϵ-entropy of a functional class, defined as follows. Let T be a domain, and let $\mathcal{F}$ be a class of functions $(f(t)\colon t \in T)$ on that domain; suppose $\mathcal{F}$ is compact for the norm $\|\cdot\|$, so that there exists an ϵ-net, i.e., a system $\mathcal{N}_\epsilon = \{f'\}$ such that

$$\sup_{f \in \mathcal{F}} \inf_{f' \in \mathcal{N}_\epsilon} \|f - f'\| \le \epsilon. \tag{5.1}$$

Let $N(\epsilon, \mathcal{F}, \|\cdot\|)$ denote the minimal cardinality of all such ϵ-nets. The Kolmogorov ϵ-entropy for $(\mathcal{F}, \|\cdot\|)$ is then

$$H_\epsilon(\mathcal{F}, \|\cdot\|) = \log_2 N(\epsilon, \mathcal{F}, \|\cdot\|). \tag{5.2}$$

It is the least number of bits required to specify any arbitrary member of $\mathcal{F}$ to within accuracy ϵ. In essence, Kolmogorov proposed a notion of data compression for *classes of functions* while Shannon's theory concerned compression for *stochastic processes*.

There are some formal similarities between the problems addressed by Shannon's $R(D)$ and Kolmogorov's H_ϵ. To make these clear, notice that in each case, we consider a "library of instances"—either a function class $\mathcal{F}$ or a stochastic process X, each case yielding as typical elements functions defined on a common domain T—and we measure approximation error by the same norm $\|\cdot\|$.

In both the Shannon and Kolmogorov theories we encode by first constructing finite lists of representative elements—in one case, the list is called a codebook; in the other case, a net. We represent an object of interest by its closest representative in the list, and we may record simply the index into our list. The length in bits of such a recording is called in the Shannon case the rate of the codebook; in the Kolmogorov case, the entropy of the net. Our goal is to minimize the number of bits while

	Shannon Theory	Kolmogorov Theory
Library	X Stochastic	$f \in \mathcal{F}$
Representers	Codebook $\mathcal{C}$	Net $\mathcal{N}$
Fidelity	$E \min_{X' \in \mathcal{C}} \|\|X - X'\|\|^2$	$\max_{f \in \mathcal{F}} \min_{f' \in \mathcal{N}} \|\|f - f'\|\|^2$
Complexity	$\log \#\mathcal{C}$	$\log \#\mathcal{N}$

achieving sufficient fidelity of reproduction. In the Shannon theory this is measured by mean discrepancy across random realizations; in the Kolmogorov theory this is measured by the maximum discrepancy across arbitrary members of $\mathcal{F}$. These comparisons may be summarized in the table at the top of this page.

In short, the two theories are parallel—except that one of the theories postulates a library of samples arrived at by sampling a stochastic process, while the other selects arbitrary elements of a functional class.

While there are intriguing parallels between the $R(D)$ and H_ϵ concepts, the two approaches have developed separately, in very different contexts. Work on $R(D)$ has mostly stayed in the original context of communication/storage of random process data, while work with H_ϵ has mostly stayed in the context of questions in mathematical analysis: the Kolmogorov entropy numbers control the boundedness of Gaussian processes [35] and the properties of certain operators [10], [36], of convex sets [78], and of statistical estimators [6], [67].

At the general level which Kolmogorov proposed, almost nothing useful can be said about the structure of an optimal ϵ-net, nor is there any principle like mutual information which could be used to derive a formal expression for the cardinality of the ϵ-net.

However, there is a particularly interesting case in which we can say more. Consider the following typical setting for ϵ-entropy. Let T be the circle $T = [0, 2\pi)$ and let $W^m_{2,0}(\gamma)$ denote the collection of all functions $f = (f(t)\colon\ t \in T)$ such that $||f||^2_{L^2(T)} + ||f^{(m)}||^2_{L^2(T)} \leq \gamma^2$. Such functions are called "differentiable in quadratic mean" or "differentiable in the sense of H. Weyl." For approximating functions of this class in L^2-norm, we have the precise asymptotics of the Kolmogorov ϵ-entropy [34]

$$H_\epsilon(W^m_{2,0}(\gamma)) \sim 2m(\log_2 e)(\gamma/2\epsilon)^{1/m}, \qquad \epsilon \to 0. \quad (5.3)$$

A transform-domain coder can achieve this H_ϵ asymptotic. One divides the frequency domain into subbands K_b, defined exactly as in (4.1) and (4.2). Then one takes the Fourier coefficients θ_k of the object f, obtaining blocks of coefficients $\theta^{(b)}$. Treating these coefficients now as if they were arbitrary members of spheres of radius $\rho_b = ||\theta^{(b)}||$, one encodes the coefficients using an ϵ_b-net for the sphere of radius ρ_b. One represents the object θ by concatenating a prefix code together with the code for the individual subbands. The prefix code records digital approximations to the (ϵ_b, ρ_b) pairs for subbands, and requires asymptotically a small number of bits. The body code simply concatenates the codes for each of the individual subbands. With the right fidelity allocation—i.e., choice of ϵ_b—the resulting code has a length described by the right side of (5.3).

VI. The JPEG Setting

We now discuss the emergence of transform coding ideas in practical coders. The discrete-time setting of practical coders makes it expedient to abandon the functional point of view throughout this section, in favor of a viewpoint based on sampled data.

A. History

Transform coding plays an important role for images and audio compression where several successful standards incorporate linear transforms. The success and wide acceptance of transform coding in practice is due to a combination of factors. The Karhunen–Loève transform and its optimality under some (restrictive) conditions form a theoretical basis for transform coding. The wide use of particular transforms like the discrete cosine transform (DCT) led to a large body of experience, including design of quantizers with human perceptual criteria. But most importantly, transform coding using a unitary matrix having a fast algorithm represents an excellent compromise in terms of computational complexity versus coding performance. That is, for a given cost (number of operations, run time, silicon area), transform coding outperforms more sophisticated schemes by a margin.

The idea of compressing stochastic processes using a linear transformation dates back to the 1950's [63], when signals originating from a vocoder were shown to be compressible by a transformation made up of the eigenvectors of the correlation matrix. This is probably the earliest use of the Karhunen–Loève transform (KLT) in data compression. Then, in 1963, Huang and Schultheiss [55] did a detailed analysis of block quantization of random variables, including bit allocation. This forms the foundation of transform coding as used in signal compression practice. The approximation of the KLT by trigonometric transforms, especially structured transforms allowing a fast algorithm, was done by a number of authors, leading to the proposal of the discrete cosine transform in 1974 [1]. The combination of discrete cosine transform, scalar quantization, and entropy coding was studied in detail for image compression, and then standardized in the late 1980's by the joint picture experts group (JPEG), leading to the JPEG image compression standard that is now widely used. In the meantime, another generation of image coders, mostly based on wavelet decompositions and elaborate quantization and entropy coding, are being considered for the next standard, called JPEG-2000.

B. The Standard Model and the Karhunen–Loève Transform

The structural facts described in Section II, concerning $R(D)$ for Gaussian random processes, become very simple in the case of Gaussian random vectors. Compression of a vector of correlated Gaussian random variables factors into a linear transform followed by independent compression of the transform coefficients.

Consider $\boldsymbol{X} = [X_0 X_1 \cdots X_{N-1}]^T$, a size N vector of zero mean random variables and $\boldsymbol{Y} = [Y_0 Y_1 \cdots Y_{N-1}]^T$ the vector of random variables after transformation by $\boldsymbol{T}$, or $\boldsymbol{Y} = \boldsymbol{T} \cdot \boldsymbol{X}$. Define $\boldsymbol{R_X} = E[\boldsymbol{X}\boldsymbol{X}^T]$ and $\boldsymbol{R_Y} = E[\boldsymbol{Y}\boldsymbol{Y}^T]$ as autocovariance matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Since $\boldsymbol{R_X}$ is symmetric and positive-semidefinite, there is a full set of orthogonal eigenvectors with nonnegative eigenvalues. The Karhunen–Loève transform matrix $\boldsymbol{T}_{\mathrm{KL}}$ is defined as the matrix of unit-norm eigenvectors of $\boldsymbol{R_X}$ ordered in terms of decreasing eigenvalues, that is,

$$\boldsymbol{R_X}\boldsymbol{T}_{\mathrm{KL}} = \boldsymbol{T}_{\mathrm{KL}}\boldsymbol{\Lambda}, \qquad \boldsymbol{\Lambda} = \mathrm{diag}\,(\lambda_0, \lambda_1, \cdots, \lambda_{N-1})$$

where $\lambda_i \geq \lambda_j \geq 0, i<j$ (for simplicity, we will assume that $\lambda_i > 0$). Clearly, transforming X with $\boldsymbol{T}_{\mathrm{KL}}^T$ will diagonalize $\boldsymbol{R_Y}$

$$\boldsymbol{R_Y} = E[\boldsymbol{T}_{\mathrm{KL}}^T \boldsymbol{X}\boldsymbol{X}^T \boldsymbol{T}_{\mathrm{KL}}] = \boldsymbol{T}_{\mathrm{KL}}^T \boldsymbol{R_X} \boldsymbol{T}_{\mathrm{KL}} = \boldsymbol{\Lambda}.$$

The KLT satisfies a best linear approximation property in the mean-squared error sense which follows from the eigenvector choices in the transform. That is, if only a fixed subset of the transform coefficients are kept, then the best transform is the KLT.

The importance of the KLT for compression comes from the following standard result from source coding [46]. A size-N Gaussian vector source $\boldsymbol{X}$ with correlation matrix $\boldsymbol{R_X}$ and mean zero is to be coded with a linear transform. Bits are allocated optimally to the transform coefficients (using reverse waterfilling). Then the transform that minimizes the MSE in the limit of fine quantization of the transform coefficients is the Karhunen–Loève transform $\boldsymbol{T}_{\mathrm{KL}}$. The coding gain due to optimal transform coding over straight PCM coding is

$$\frac{D_{\mathrm{PCM}}}{D_{\mathrm{KLT}}} = \frac{\sigma_x^2}{\left(\prod_{i=0}^{N-1} \sigma_i^2\right)^{1/N}} = \frac{1/N \sum_{i=0}^{N-1} \sigma_i^2}{\left(\prod_{i=0}^{N-1} \sigma_i^2\right)^{1/N}} \tag{6.1}$$

where we used $N \cdot \sigma_x^2 = \Sigma\, \sigma_i^2$. Recalling that the variances σ_i^2 are the eigenvalues of $\boldsymbol{R_X}$, it follows that the coding gain is the ratio of the arithmetic and geometric means of the eigenvalues of the autocorrelation matrix.

Using reverse waterfilling, the above construction can be used to derive the $R(D)$ function of i.i.d. Gaussian vectors [19]. However, an important point is that in a practical setting and for complexity reasons, only scalar quantization is used on the transform coefficients (see Fig. 1). The high rate scalar distortion rate function (with entropy coding) for i.i.d. Gaussian samples of variance σ^2 is given by $D_s(R) = (\pi e)/6 \cdot \sigma^2 \cdot 2^{-2R}$ while the Shannon distortion rate function is $D(R) = \sigma^2 \cdot 2^{-2R}$ (using block coding). This means that a penalty of about a quarter bit per sample is paid, a small price at high rates or small distortions.

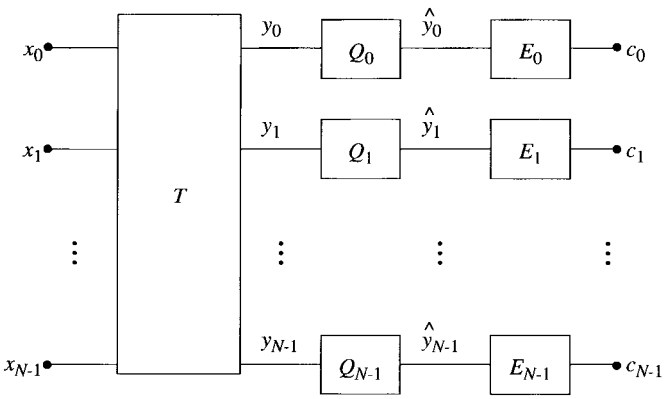

Fig. 1. Transform coding system, where T is a unitary transform, Q_i are scalar quantizers, and E_i are entropy coders.

C. The Discrete Cosine Transform

To make the KLT approach to block coding operational requires additional steps. Two problems need to be addressed: the signal dependence of the KLT (finding eigenvectors of the correlation matrix), and the complexity of computing the KLT (N^2 operations). Thus fast fixed transforms (with about $N \log N$ operations) leading to approximate diagonalization of correlation matrices are used. The most popular among these transforms is the discrete cosine transform, which has the property that it diagonalizes approximately the correlation matrix of a first-order Gauss–Markov process with high correlation ($\rho \to 1$), and also the correlation matrix of an arbitrary Gauss–Markov process (with correlation of sufficient decay, $\Sigma_{k=0}^{\infty} kr^2(k) < \infty$) and block sizes $N \to \infty$. The DCT is closely related to the discrete Fourier transform, and thus can be computed with a fast Fourier transform like algorithm in $N \log N$ operations. This is a key issue: the DCT achieves a good compromise between coding gain or compression, and computational complexity. Therefore, for a given computational budget, it can actually outperform the KLT [49].

VII. The Common Gaussian Model

At this point we have looked at three different settings in which we can interpret the phrase "data compression." In each case *we have available a library of instances which we would like to represent with few bits.*

- a) In Section II (on $R(D)$ theory) we are considering the instances to be realizations of Gaussian processes. The library is the collection of all such realizations.
- b) In Section V (on ϵ-entropy) we are considering the instances to be smooth functions. The library $\mathcal{F}$ is the collection of such smooth functions obeying the constraint $||f||_{L^2(T)}^2 + ||f^{(m)}||_{L^2(T)}^2 \leq \gamma^2$.

- c) In Section VI (on JPEG), we are considering the instances to be existing or future digital images. The library is implicitly the collection of all images of potential interest to the JPEG user population.

We can see a clear similarity in the coding strategy used in each setting.

- Transform into the frequency domain.
- Break the transform into homogeneous subbands.
- Apply simple coding schemes to the subbands.

Why the common strategy of transform coding?

The theoretically tightest motivation for transform coding comes from Shannon's $R(D)$ theory, which tells us that in order to best encode a Gaussian process, one should transform the process realization to the Karhunen–Loève domain, where the resulting coordinates are independent random variables. This sequence of independent Gaussian variables can be coded by traditional schemes for coding discrete memoryless sources.

So when we use transform coding in another setting—ϵ-entropy or JPEG—it appears that we are *behaving as if that setting could be modeled by a Gaussian process*.

In fact, it is sometimes said that the JPEG scheme is appropriate for real image data because if image data were first-order Gauss–Markov, then the DCT would be approximately the Karhunen–Loève transform, and so JPEG would be approximately following the script of the $R(D)$ story. Implicitly, the next statement is "and real image data behave something like first-order Gauss–Markov."

What about ϵ-entropy? In that setting there is no obvious "randomness," so it would seem unclear how a connection with Gaussian processes could arise. In fact, a proof of (5.3) can be developed by exhibiting just such a connection [34]; one can show that there are Gaussian random functions whose sample realizations obey, with high probability, the constraint $||f||^2_{L^2(T)} + ||f^{(m)}||^2_{L^2(T)} \leq \gamma^2$ and for which the $R(D)$ theory of Shannon accurately matches the number of bits required, in the Kolmogorov theory, to represent f within a distortion level ϵ^2 (i.e., the right side of (5.3)). The Gaussian process with this property is a process naturally associated with the class $\mathcal{F}$ obeying the indicated smoothness constraint—the least favorable process for Shannon data-compression; a successful Kolmogorov-net for the process $\mathcal{F}$ will be effectively a successful Shannon codebook for the least favorable process. So even in the Kolmogorov case, transform coding can be motivated by recourse to $R(D)$ theory for Gaussian processes, and to the idea that the situation can be modeled as a Gaussian one. (That a method derived from Gaussian assumptions helps us in other cases may seem curious. This is linked to the fact that the Gaussian appears as a worst case scenario. Handling the worst case well will often lead to adequate if not optimal performance for more favorable cases.)

VIII. Getting the Model Right

We have so far considered only a few settings in which data compression could be of interest. In the context of $R(D)$ theory, we could be interested in complex non-Gaussian processes; in H_ϵ theory we could be interested in functional classes defined by norms other than those based on L^2; in image coding we could be interested in particular image compression tasks, say specialized to medical imagery, or to satellite imagery.

Certainly, the simple idea of Fourier transform followed by block i.i.d. Gaussian coding cannot be universally appropriate. *As the assumptions about the collection of instances to be represented change*, presumably the corresponding *optimal representation will change*. Hence it is important to explore a range of modeling assumptions and to attempt to get the assumptions right! Although Shannon's ideas have been very important in supporting diffusion of frequency-domain coding in practical lossy compression, we feel that he himself would have been the first to suggest a careful examination of the underlying assumptions, and to urge the formulation of better assumptions. (See, for instance, his adhortations in [85].)

In this section we consider a wider range of models for the libraries of instances to be compressed, and see how alternative representations emerge as useful.

A. Some Non-Gaussian Models

Over the last decade, studies of the statistics of natural images have repeatedly shown the non-Gaussian character of image data. While images make up only one application area for data compression, the evidence is quite interesting.

Empirical studies of wavelet transforms of images, considering histograms of coefficients falling in a common subband, have uncovered markedly non-Gaussian structure. As noted by many people, subband histograms are consistent with probability densities having the form $C \cdot \exp\{-|u|^\mu\}$, where the exponent "μ" would be "2" if the Gaussian case applied, but where one finds radically different values of "μ" in practice; e.g., Simoncelli [87] reports evidence for $\mu = 0.7$. In fact, such generalized Gaussian models have been long used to model subband coefficients in the compression literature (e.g., [101]). Field [37] investigated the fourth-order cumulant structure of images and showed that it was significantly nonzero. This is far out of line with the Gaussian model, in which all cumulants of order three and higher vanish.

In later work, Field [38] proposed that wavelet transforms of images offered probability distributions which were "sparse." A simple probability density with such a sparse character is the Gaussian scale mixture $(1-\epsilon)\phi(x/\delta)/\delta + \epsilon\phi(x)$, where ϵ and δ are both small positive numbers; this corresponds to data being of one of two "types:" "small," the vast majority, and "large," the remaining few. It is not hard to understand where the two types come from: a wavelet coefficient can be localized to a small region which contains an edge, or which does not contain an edge. If there is no edge in the region, it will be "small;" if there is an edge, it will be "large."

Stationary Gaussian models are very limited and are unable to duplicate these empirical phenomena. Images are best thought of as spatially stationary stochastic processes, since logically the position of an object in an image is rather arbitrary, and a shift of that object to another position would

produce another equally valid image. But if we impose stationarity on a Gaussian process we cannot really exhibit both edges and smooth areas. A stationary Gaussian process must exhibit a great deal of spatial homogeneity. From results in the mean-square calculus we know that if such a process is mean-square-continuous at a point, it is mean-square-continuous at every point. Clearly, most images will not fit this model adequately.

Conditionally Gaussian models offer an attractive way to maintain ties with the Gaussian case while exhibiting globally non-Gaussian behavior. In such models, image formation takes place in two stages. An initial random experiment lays down regions separated by edges, and then in a subsequent stage each region is assigned a Gaussian random field.

Consider a simple model of random piecewise-smooth functions in dimension one, where the piece boundaries are thrown down at random, say by a Poisson process, the pieces are realizations of (different) Gaussian processes (possibly stationary), and discontinuities are allowed across the boundaries of the pieces [12]. This simple one-dimensional model can replicate some of the known empirical structure of images, particularly the sparse histogram structure of wavelet subbands and the nonzero fourth-order cumulant structure.

B. Adaptation, Resource Allocation, and Nonlinearity

Unfortunately, when we leave the domain of Gaussian models, we lose the ability to compute $R(D)$ in such great generality. Instead, we begin to operate heuristically. Suppose, for example, we employ a conditionally Gaussian model. There is no general solution for $R(D)$ for such a class; but it seems reasonable that the two-stage structure of the model gives clues about optimal coding; accordingly, one might suppose that an effective coder should factor into a part that adapts to the apparent segmentation structure of the image and a part that acts in a traditional way conditional on that structure. In the simple model of piecewise-smooth functions in dimension one, it is clear that coding in long blocks is useful for the pieces, and that the coding must be adapted to the characteristics of the pieces. However, discontinuities must be well-represented also. So it seems natural that one attempts to identify an empirically accurate segmentation and then adaptively code the pieces. If transform coding ideas are useful in this setting, it might seem that they would play a role subordinate to the partitioning—i.e., appearing only in the coding of individual pieces. It might seem that applying a single global orthogonal transform to the data is simply not compatible with the assumed two-stage structure.

Actually, transform coding *is* able to offer a degree of adaptation to the presence of a segmentation. The wavelet transform of an object with discontinuities will exhibit large coefficients in the neighborhood of discontinuities, and, at finer scales, will exhibit small coefficients away from discontinuities. If one designs a coder which does well in representing such "sparse" coefficient sequences, it will attempt to represent all the coefficients at coarser scales, while allocating bits to represent only those few big coefficients at finer scales. Implicitly, coefficients at coarser scales represent the structure of pieces, and coefficients at finer scales represent discontinuities between the pieces. The resource allocation is therefore achieving some of the same effect as an explicit two-stage approach.

Hence, adaptivity to the segmentation can come from applying a fixed orthogonal transform together with adaptive resource allocation of the coder. Practical coding experience supports this. Traditional transform coding of i.i.d. Gaussian random vectors at high rate assumes a fixed rate allocation per symbol, but practical coders, because they work at low rate and use entropy coding, typically adapt the coding rate to the characteristics of each block. Specific adaptation mechanisms, using context modeling and implicit or explicit side-information are also possible.

Adaptive resource allocation with a fixed orthogonal transform is closely connected with a mathematical procedure which we will explore at length in Sections XIV and XV: nonlinear approximation using a fixed orthogonal basis. Suppose that we have an orthogonal basis and we wish to approximate an object using only n basis functions. In traditional linear approximation, we would consider using the first-n basis functions to form such an approximation. In nonlinear approximation, we would consider using the best-n basis functions, i.e., to adaptively select the n terms which offer the best approximation to the particular object being considered. This adaptation is a form of resource allocation, where the resources are the n terms to be used. Because of this connection, we will begin to refer to "the nonlinear nature of the approximation process" offered by practical coders.

C. Variations on Stochastic Process Models

To bring home the remarks of the last two subsections, we consider some specific variations on the stochastic process models of Section II. In these variations, we will consider processes that are non-Gaussian; and we will compare useful coding strategies for those processes with the coding strategies for the Gaussian processes having the same second-order statistics.

- *Spike Process*. For this example, we briefly leave the functional viewpoint.

  Consider the following simple discrete-time random process, generating a single "spike." Let $x(n) = \alpha \cdot \delta(n-k)$ where $n, k \in [0, \cdots, N-1], k$ is uniformly distributed between 0 and $N-1$ and α is $N(0, \sigma^2)$. That is, after picking a random location k, one puts a Gaussian random variable at that location. The autocorrelation $\boldsymbol{R_X}$ is equal to $(\sigma^2/N) \cdot \boldsymbol{I}$, thus the KLT is the identity transformation. Allocating R/N bits to each coefficient leads to a distortion of order $2^{-2(R/N)}$ for the single nonzero coefficient. Hence the distortion-rate function describing the operational performance of the Gaussian codebook coder in the KLT domain has

  $$D_{\mathrm{KL}}(R) \approx c \cdot \sigma^2 \cdot 2^{-2(R/N)}.$$

  Here the constant c depends on the quantization and coding of the transform coefficients.

  An obvious alternate scheme at high rates is to spend $\log_2(N)$ bits to address the nonzero coefficient, and use

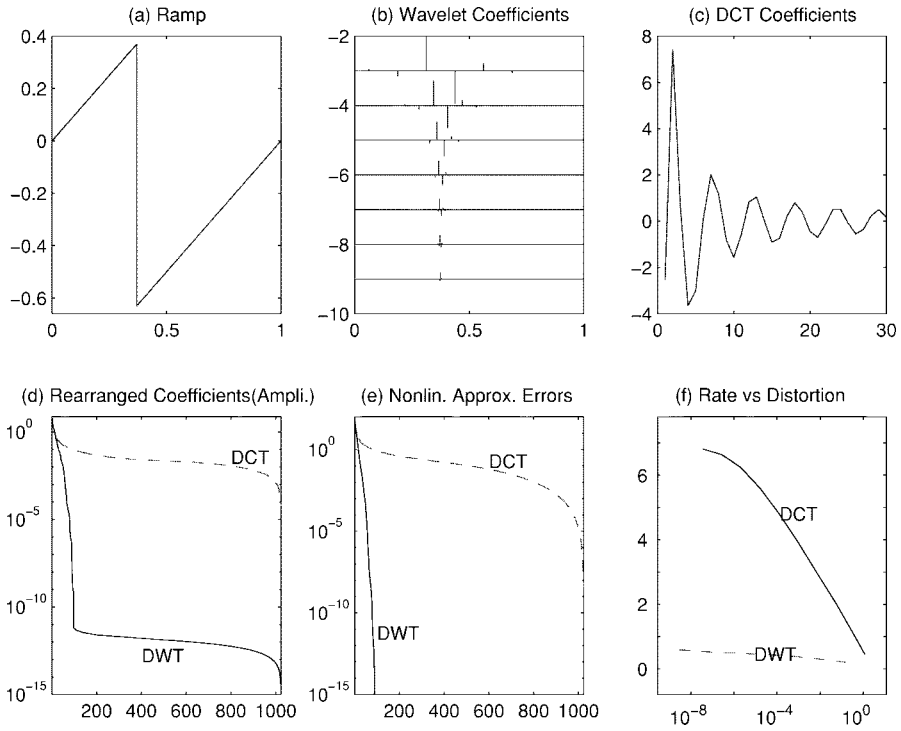

Fig. 2. (a) Realization of *Ramp*. (b) Wavelet and (c) DCT coefficients. (d) Rearranged coefficients. (e) Nonlinear approximation errors (14.2). (f) Operating performance curves of scalar quantization coding.

the remaining $R-\log_2(N)$ bits to represent the Gaussian variable. This *position-indexing* method leads to

$$D_p(R) \approx c \cdot \sigma^2 \cdot 2^{-2(R-\log_2(N))}.$$

This coder relies heavily on the non-Gaussian character of the joint distribution of the entries in $x(n)$, and for $R \gg \log_2(N)$ this non-Gaussian coding clearly outperforms the former, Gaussian approximation method. While this is clearly a very artificial example, it makes the point that if location (or phase) is critical, then time-invariant, linear methods like the KLT followed by independent coding of the transform coefficients are suboptimal.

Images are very phase critical: edges are among the most visually significant features of images, and thus efficient position coding is of essence. So it comes as no surprise that some nonlinear approximation ideas made it into standards, namely, ideas where addressing of large coefficients is efficiently solved.

- *Ramp Process*. Yves Meyer [75] proposed the following model. We have a process $X(t)$ defined on $[0,1]$ through a single random variable τ uniformly distributed on $[0,1]$ by

$$X(t) = t - 1_{\{t \geq \tau\}}.$$

This is a very simple process, and very easy to code accurately. A reasonable coding scheme would be to extract τ by locating the jump of the process and then quantizing it to the required fidelity.

On the other hand, *Ramp* is covariance equivalent to the Brownian Bridge process $B_0(t)$ which we mentioned already in Section II, the Gaussian zero-mean process on $[0,1]$ with covariance $\mathrm{Cov}(B_0(t), B_0(s)) = \min(t,s) - st$. An asymptotically $R(D)$-optimal approach to coding Brownian Bridge can be based on the Karhunen–Loève transform; as we have seen, in this case the sine transform. One takes the sine transform of the realization, breaks the sequence of coefficients into subbands obeying (4.1) and (4.2) and then treats the coefficients, in subbands, exactly as in discrete memoryless source compression.

Suppose we ignored the non-Gaussian character of the *Ramp* process and simply applied the same coder we would use for Brownian Bridge. After all, the two are covariance-equivalent. This would result in orders of magnitude more bits than necessary. The coefficients in the sine transform of *Ramp* are random; their typical size is measured in mean square by the eigenvalues of the covariance—namely, $\lambda_k = (4\pi^2 k^2)^{-1}$. In order to accurately represent the Ramp process with distortion D, we must code the first $\#K(D) \asymp D^{-1}$ coefficients, at rates exceeding 1 bit per coefficient. How many coefficients does it take to represent a typical realization of *Ramp* with a relative error of 1%? About 10^5.

On the other hand, as Meyer pointed out, the wavelet coefficients of *Ramp* decay very rapidly, essentially exponentially. As a result, very simple scalar quantization schemes based on wavelet coefficients can capture realizations of *Ramp* with 1% accuracy using a few dozens rather than tens of thousands of coefficients, and with a corresponding advantage at the level of bits; this is illustrated in Fig. 2.

The point here is that if we pay attention to second-order statistics only, and adopt an approach that would be good under a Gaussian model, we may pay orders of magnitude more bits than would be necessary for

coding the process under a more appropriate model. By abandoning the Karhunen–Loève transform in this non-Gaussian case, we get a transform in which very simple scalar quantization works very well.

Note that we could in principle build a near-optimal scheme by transform coding with a coder based on Fourier coefficients, but we would have to apply a much more complex quantizer; it would have to be a vector quantizer. (Owing to Littlewood–Paley theory described later, it is possible to say what the quantizer would look like; it would involve quantizing coefficients near wavenumber k in blocks of size roughly $k/2$. This is computationally impractical.)

D. Variations on Function Class Models

In Section V, we saw that subband coding of Fourier coefficients offered an essentially optimal method, under the Kolmogorov ϵ-entropy model, of coding objects f known *a priori* to obey L^2 smoothness constraints $||f||^2_{L^2(T)}+||f^{(m)}||^2_{L^2(T)} \leq \gamma^2$.

While this may not be apparent to outsiders, there are major differences in the implications of various smoothness constraints. Suppose we maintain the L^2 distortion measure, but make the seemingly minor change from the L^2 form of constraint to an L^p form, $||f||^p_{L^p(T)}+||f^{(m)}||^p_{L^p(T)} \leq \gamma^p$ with $p<2$. This can cause major changes in what constitutes an underlying optimal strategy. Rather than transform coding in the frequency domain, we can find that transform coding in the wavelet domain is appropriate.

Bounded Variation Model: As a simple example, consider the model that the object under consideration is a function $f(t)$ of a single variable that is of bounded variation. Such functions f can be interpreted as having derivatives which are signed measures, and then we measure the norm by

$$||f||_{\mathrm{BV}} = \int |df|.$$

The important point is such f can have jump discontinuities, as long as the sum of the jumps is finite. Hence, the class of functions of bounded variation can be viewed as a model for functions which have discontinuities; for example, a scan line in a digital image can be modeled as a typical BV function.

An interesting fact about BV functions is that they can be essentially characterized by their Haar coefficients. The BV functions with norm $\leq\gamma$ obey an inequality

$$\sup_j \sum_k |\alpha_{j,k}| 2^{j/2} \leq 4\gamma$$

where $\alpha_{j,k}$ are the Haar wavelet expansion coefficients. It is almost the case that every function that obeys this constraint is a BV function. This says that geometrically, the class of BV functions with norm $\leq\gamma$ is a convex set inscribed in a family of ℓ^1 balls.

An easy coder for functions of Bounded Variation can be based on scalar quantization of Haar coefficients. However, scalar quantization of the Fourier coefficients would not work nearly as well; as the desired distortion $\epsilon \to 0$, the number of bits for Fourier/scalar quantization coding can be orders of magnitude worse than the number of bits for wavelet/scalar quantization coding. This follows from results in Sections XV and XVI below.

E. Variations on Transform Coding and JPEG

When we consider transform coding as applied to empirical data, we typically find that a number of simple variations can lead to significant improvements over what the strict Gaussian $R(D)$ theory would predict. In particular, we see that when going from theory to practice, KLT as implemented in JPEG becomes nonlinear approximation!

The image is first subdivided into blocks of size N by N (N is typically equal to 8 or 16) and these blocks are treated independently. Note that blocking the image into independent pieces allows to adapt the compression to each block individually. An orthonormal basis for the two-dimensional blocks is derived as a product basis from the one-dimensional DCT. While not necessarily best, this is an efficient way to generate a two-dimensional basis.

Now, quantization and entropy coding is done in a manner that is quite at variance with the classical setup. First, based on perceptual criteria, the transform coefficient $y(k,l)$ is quantized with a uniform quantizer of stepsize $\Delta_{k,l}$. Typically, $\Delta_{k,l}$ is small for low frequencies, and large for high ones, and these stepsizes are stored in a quantization matrix $\boldsymbol{M}_Q$. Technically, one could pick different quantization matrices for different blocks in order to adapt, but usually, only a single scale factor α is used to multiply $\boldsymbol{M}_Q$, and this scale factor can be adapted depending on the statistics in the block. Thus the approximate representation of the (k,l)th coefficient is $\hat{y}(k,l) = Q[y(k,l), \alpha\Delta_{k,l}]$ where $Q[y,\Delta] = \Delta\cdot\lfloor y/\Delta \rfloor + \Delta/2$. The quantized variable $\hat{y}(k,l)$ is discrete with a finite number of possible values ($y(k,l)$ is bounded) and is entropy-coded.

Since there is no natural ordering of the two-dimensional DCT plane, yet known efficient entropy coding techniques work on one-dimensional sequences of coefficients, a prescribed 2D to 1D scanning is used. This so-called "zig-zag" scan traverses the DCT frequency plane diagonally from low to high frequencies. For this resulting one-dimensional length-N^2 sequence, nonzero coefficients are entropy-coded, and stretches of zero coefficients are encoded using entropy coding of run lengths. An *end-of-block* (EOB) symbol terminates a sequence of DCT coefficients when only zeros are left (which is likely to arrive early in the sequence when coarse quantization is used).

Let us consider two extreme modes of operation: In the first case, assume very fine quantization. Then, many coefficients will be nonzero, and the behavior of the rate–distortion tradeoff is dominated by the quantization and entropy coding of the individual coefficients, that is, $D(R) \sim 2^{-2R}$. This mode is also typical for high variance regions, like textures.

In the second case, assume very coarse quantization. Then, many coefficients will be zero, and the run-length coding is an efficient indexing of the few nonzero coefficients. We are

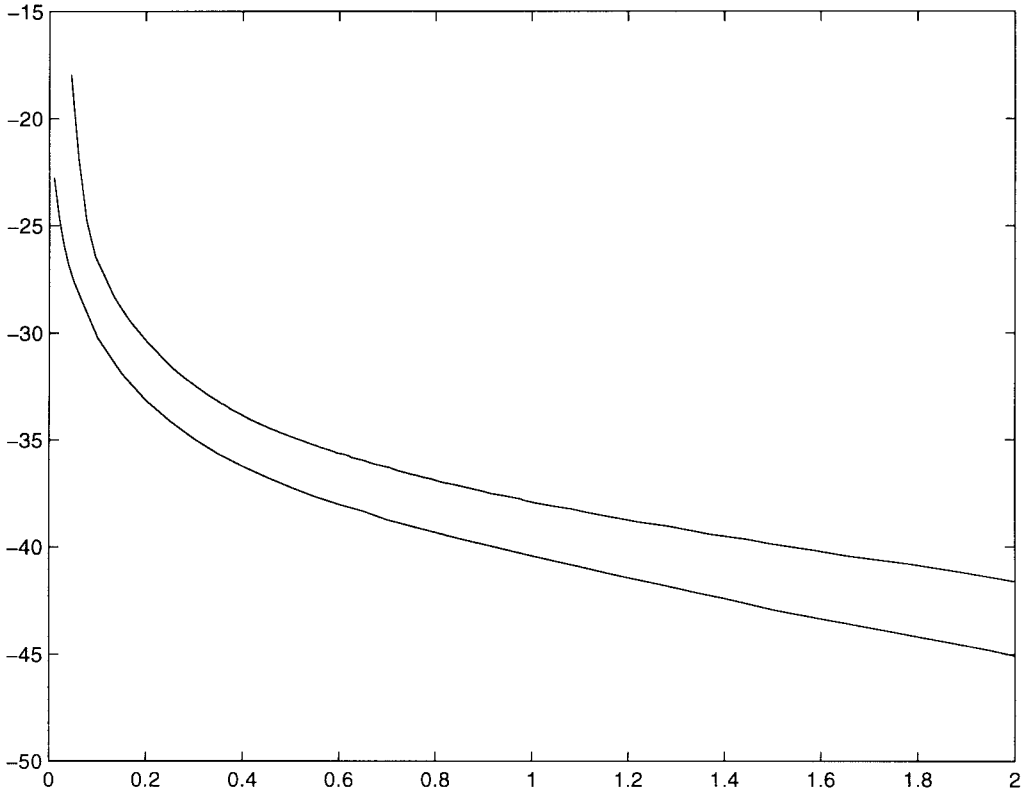

Fig. 3. Performance of real transform coding systems. The logarithm of the MSE is shown for JPEG (top) and SPIHT (bottom). Above about 0.5 bit/pixel, there is the typical -6 dB per bit slope, while at very low bit rate, a much steeper slope is achieved.

in a nonlinear approximation case, since the image block is approximated with a few basis vectors corresponding to large inner products. Then, the $D(R)$ behavior is very different, dominated by the faster decay of ordered transform coefficients, which in turn is related to the smoothness class of the images. Such a behavior is also typical for structured regions, like smooth surfaces cut by edges, since the DCT coefficients will be sparse.

These two different behaviors can be seen in Fig. 3, where the logarithm of the distortion versus the bit rate per pixel is shown. The $-2R$ slope above about 0.5 bit/pixel is clear, as is the steeper slope below. An analysis of the low-rate behavior of transform codes has been done recently by Mallat and Falzon [70]; see also related work in Cohen, Daubechies, Guleryuz, and Orchard [14].

IX. Good Models for Natural Data?

We have now seen, by considering a range of different intellectual models for the class of objects of interest to us, that depending on the model we adopt, we can arrive at very different conclusions about the "best" way to represent or compress those objects. We have also seen that what seems like a good method in one model can be a relatively poor method according to another model. We have also seen that existing models used in data compression are relatively poor descriptions of the phenomena we see in natural data. We think that we may still be far away from achieving an optimal representation of such data.

A. How Many Bits for Mona Lisa?

A classic question, somewhat tongue in cheek, is: how many bits do we need to describe Mona Lisa? JPEG uses 187 Kbytes in one version. From many points of view, this is far more than the number intrinsically required.

Humans will recognize a version based on a few hundred bits. An early experiment by L. Harmon of Bell Laboratories shows a recognizable Abraham Lincoln at 756 bits, a trick also used by S. Dali in his painting "*Slave Market with Invisible Bust of Voltaire*."

Another way to estimate the number of bits in a representation is to consider an index of every photograph ever taken in the history of mankind. With a generous estimate of 100 billion pictures a year, the 100 years of photography need an index of about 44 bits. Another possibility yet is to index all pictures that can possibly be viewed by all humans. Given the world population, and the fact that at most 25 pictures a second are recognizable, a hundred years of viewing is indexed in about 69 bits.

Given that the Mona Lisa is a very famous painting, it is clear that probably a few bits will be enough (with the obvious variable length code: [is it Lena?, is it Mona Lisa?, etc. . . .]). Another approach is the interactive search of the image, for example, on the Web. A search engine prompted with a few key words will quickly come back with the answer at the top of the following page, and just a few bytes have been exchanged.

These numbers are all very suggestive when we consider estimates of the information rate of the human visual system.

http://www.paris.org/Musees/Louvre/Treasures/gifs/Mona_Lisa.jpg

Barlow [4] summarizes evidence that the many layers of processing in the human visual system reduce the information flow from several megabits per second at the retina to about 40 bits per second deep in the visual pathway.

From all points of view, *images ought to be far more compressible than current compression standards allow*.

B. The Grand Challenge

An effort to do far better compression leads to the Grand Challenge, items GC1–GC3 of Section I. However, to address this challenge by orthodox application of the Shannon theory seems to us hopeless. To understand why, we make three observations.

- *Intrinsic Complexity of Natural Data Sources:* An accurate model for empirical phenomena would be of potentially overwhelming complexity. In effect, images, or sounds, even in a restricted area of application like medical imagery, are naturally infinite-dimensional phenomena. They take place in a continuum, and in principle the recording of a sound or image cannot be constrained in advance by a finite number of parameters. The true underlying mechanism is in many cases markedly non-Gaussian, and highly nonstationary.
- *Difficulty of Characterization:* There exists at the moment no reasonable "mechanical" way to characterize the structure of such complex phenomena. In the zero-mean Gaussian case, all behavior can be deduced from properties of the countable sequence of eigenvalues of the covariance kernel. Outside of the Gaussian case, very little is known about characterizing infinite-dimensional probability distributions which would be immediately helpful in modeling real-world phenomena such as images and sounds. Instead, we must live by our wits.
- *Complexity of Optimization Problem:* If we take Shannon literally, and apply the abstract $R(D)$ principle, determining the best way to code a naturally occurring source of data would require to solve a mutual information problem involving probability distributions defined on an infinite-dimensional space. Unfortunately, it is not clear that one can obtain a clear intellectual description of such probability distributions in a form which would be manageable for actually stating the problem coherently, much less solving it.

In effect, uncovering the optimal codebook structure of naturally occurring data involves more challenging empirical questions than any that have ever been solved in empirical work in the mathematical sciences. Typical empirical questions that have been adequately solved in scientific work to date involve finding structure of very simple low-dimensional, well-constrained probability distributions.

The problem of determining the solution of the $R(D, X)$-problem, given a limited number of realizations of X, could be considered a branch of what is becoming known in statistics as "functional data analysis"—the analysis of data when the observations are images, sounds, or other functions, and so naturally viewed as infinite-dimensional. Work in that field aims to determine structural properties of the probability distribution of functional data—for example, the covariance and/or its eigenfunctions, or the discriminant function for testing between two populations. Functional data analysis has shown that many challenging issues impede the extension of simple multivariate statistical methods to the functional case [79]. Certain simple multivariate procedures have been extended to the functional case: principal components, discriminant analysis, canonical correlations being carefully studied examples. The problem that must be faced in such work is that one has always a finite number of realizations from which one is to infer aspects of the infinite-dimensional probabilistic generating mechanism. This is a kind of rank deficiency of the data set which means that, for example, one cannot hope to get quantitatively accurate estimates of eigenfunctions of the covariance.

The mutual information optimization problem in Shannon's $R(D)$ in the general non-Gaussian case requires far more than just knowledge of a covariance or its eigenfunctions; it involves in principle all the joint distributional structure of the process. It is totally unclear how to deal with the issues that would crop up in such a generalization.

X. A Role for Harmonic Analysis

In this section, we comment on some interesting insights that harmonic analysis has to offer against the background of this "Grand Challenge."

A. Terminology

The phrase "harmonic analysis" means many things to many different people. To some, it is associated with an abstract procedure in group theory—unitary group representations [54]; to others it is associated with classical mathematical physics—expansions in special functions related to certain differential operators; to others it is associated with "hard" analysis in its modern form [90].

The usual senses of the phrase all have roots in the bold assertions of Fourier that a) "any" function can be expanded in a series of sines and cosines and that, b) one could understand the complex operator of heat propagation by understanding merely its action on certain "elementary" initial distributions—namely, initial temperature distributions following sinusoidal profiles. As is now well known, making sense in one way or another of Fourier's assertions has spawned an amazing array of concepts over the last two centuries; the theories of the Lebesgue integral, of Hilbert spaces, of L^p

spaces, of generalized functions, of differential equations, are all bound up in some way in the development, justification, and refinement of Fourier's initial ideas. So no single writer can possibly mean "*all* of harmonic analysis" when using the term "harmonic analysis."

For the purposes of this paper, harmonic analysis refers instead to a series of ideas that have evolved throughout this century, a set of ideas involving two streams of thought.

On the one hand, to develop ways to analyze functions using decompositions built through a geometrically motivated "cutting and pasting" of time, frequency, and related domains into "cells" and the construction of systems of "atoms" "associated naturally" to those "cells."

On the other hand, to use these decompositions to find characterizations, to within notions of equivalence, of interesting function classes.

It is perhaps surprising that one can combine these two ideas. *A priori* difficult to understand classes of functions in a functional space turn out to have a characterization as a superposition of "atoms" of a more or less concrete form. Of course, the functional spaces where this can be true are quite special; the miracle is that it can be done at all.

This is a body of work that has grown up slowly over the last ninety years, by patient accumulation of a set of tools and cultural attitudes little known to outsiders. As we stand at the end of the century, we can say that this body of work shows that there are many interesting questions about infinite-dimensional function classes where experience has shown that it is often difficult or impossible to obtain exact results, but where fruitful analogies do suggest similar problems which are "good enough" to enable approximate, or asymptotic solutions.

In a brief survey paper, we can only superficially mention a few of the decomposition ideas that have been proposed and a few of the achievements and cultural attitudes that have resulted; we will do so in Sections XII, XVII, and XVIII below. The reader will find that [58] provides a wealth of helpful background material complementing the present paper.

B. Relevance to the Grand Challenge

Much of harmonic analysis in this century can be characterized as carrying out a three-part program

HA1 Identify an interesting class of mathematically defined objects (functions, operators, etc.).

HA2 Develop tools to characterize the class of objects in terms of functionals derivable from an analysis of the objects themselves.

HA3 Improve the characterization tools themselves, refining and streamlining them.

This program, while perhaps obscure to outsiders, has borne interesting fruit. As we will describe below, wavelet transforms arose from a long series of investigations into the structure of classes of functions defined by L^p constraints, $p \neq 2$. The original question was to characterize function classes $\{f: \int |f|^p < \infty\}$ by analytic means, such as by the properties of the coefficients of the considered functions in an orthogonal expansion. It was found that the obvious first choice—Fourier coefficients—could not offer such a characterization when $p \neq 2$, and eventually, after a long series of alternate characterizations were discovered, it was proved that wavelet expansions offered such characterizations—i.e., that by looking at the wavelet coefficients of a function, one could learn the L^p-norm to within constants of equivalence, $1 < p < \infty$. It was also learned that for $p \in \{1, \infty\}$ no norm characterization was possible, but after replacing L^1 by the closely related space H^1 and L^∞ by the closely related space BMO (the space of functions of Bounded Mean Oscillation), the wavelet coefficients again contained the required information for knowing the norm to within constants of equivalence.

The three parts HA1–HA3 of the harmonic analysis program are entirely analogous to the three steps GC1–GC3 in the Grand Challenge for data compression—except that for data compression, the challenge is to deal with *naturally occurring data sources* while for harmonic analysis the challenge is to deal with *mathematically interesting classes of objects*.

It is very striking to us that the natural development of harmonic analysis in this century, while in intention completely unrelated to problems of data compression, has borne fruit which seems very relevant to the needs of the data compression community. Typical byproducts of this effort so far include the fast wavelet transform, and lossy wavelet domain coders which exploit the "tree" organization of the wavelet transform. Furthermore, we are aware of many other developments in harmonic analysis which have not yet borne fruit of direct impact on data compression, but seem likely to have an impact in the future.

There are other insights available from developments in harmonic analysis. In the comparison between the three-part challenge facing data compression and the three-part program of harmonic analysis, the messiness of understanding a natural data source—which requires dealing with specific phenomena in all their particularity—is replaced by the precision of understanding a class with a formal mathematical definition. Thus harmonic analysis operates in a more ideal setting for making intellectual progress; but sometimes progress is not as complete as one would like. It is accepted by now that many characterization problems of function classes cannot be exactly solved. Harmonic analysis has shown that often one can make substantial progress by replacing hard characterization problems with less demanding problems where answers can be obtained explicitly. It has also shown that such cruder approximations are still quite useful and important.

A typical example is the study of operators. The eigenfunctions of an operator are fragile, and can change radically if the operator is only slightly perturbed in certain ways. It is in general difficult to get explicit representations of the eigenfunctions, and to compute them. Harmonic analysis, however, shows that for certain problems, we can work with "almost eigenfunctions" that "almost diagonalize" operators. For example, wavelets work well on a broad range of operators, and moreover, they lend themselves to concrete fast computational algorithms. That is, the exact problem, which is potentially intractable, is replaced by an approximate problem for which computational solutions exist.

C. Survey of the Field

In the remainder of this paper we will discuss a set of developments in the harmonic analysis community and how they can be connected to results about data compression. We think it will become clear to the reader that in fact the lessons that have been learned from the activity of harmonic analysis are very relevant to facing the Grand Challenge for data compression.

XI. Norm Equivalence Problems

The problem of characterizing a class of functions $F = \{f: ||f||_F < \infty\}$, where $||\cdot||_F$ is a norm of interest, has occupied a great deal of attention of harmonic analysts in this century. The basic idea is to relate the norm, defined in say continuum form by an integral, to an equivalent norm defined in discrete form

$$||f||_F \asymp ||\theta(f)||_{\boldsymbol{f}}. \tag{11.1}$$

Here $\theta = \theta(f)$ denotes a collection of coefficients arising in a decomposition of f; for example, these coefficients would be obtained by

$$\theta_k = \langle f, \phi_k \rangle$$

if the (ϕ_k) made up an orthonormal basis. The norm $||\cdot||_F$ is a norm defined on the continuum object f and the norm $||\cdot||_{\boldsymbol{f}}$ is a norm defined on the corresponding discrete object $\theta(f)$. The equivalence symbol $\asymp$ in (11.1) means that there are constants A and B, not depending on f, so that

$$A||f||_F \le ||\theta(f)||_{\boldsymbol{f}} \le B||f||_F. \tag{11.2}$$

The significance is that the coefficients θ contain within them the information needed to approximately infer the size of f in the norm $||\cdot||_F$. One would of course usually prefer to have $A = B$, in which case the coefficients characterize the size of f precisely, but often this kind of *tight* characterization is beyond reach.

The most well-known and also tightest form of such a relationship is the Parseval formula, valid for the continuum norm of L^2: $||f||_{L^2} = (\int_T |f(t)|^2\,dt)^{1/2}$. If the $(\phi_k)_k$ constitute a complete orthonormal system for $L^2(T)$, then we have

$$||f||_{L^2(T)} = \left(\sum_k |\theta_k|^2\right)^{1/2}. \tag{11.3}$$

Another frequently encountered relationship of this kind is valid for functions on the circle $T = [0, 2\pi)$, with (ϕ_k) the Fourier basis of Section II. Then we have a norm equivalence for a norm defined on the mth derivative

$$||f^{(m)}||_{L^2(T)} = \left(\sum_k k^{2m}(|\theta_{2k}|^2 + |\theta_{2k+1}|^2)\right)^{1/2}. \tag{11.4}$$

These two equivalences are, of course, widely used in the mathematical sciences. They are beautiful, but also potentially misleading. A naive reading of these results might promote the expectation that one can frequently have tightness $A = B = 1$ in characterization results, or that the Fourier basis works in other settings as well.

We mention now five kinds of norms for which we might like to solve norm-equivalence, and for which the answers to all these norm-equivalence problems are by now well-understood.

- *L^p-norms:* Let $1 \le p < \infty$. The L^p-norm is, as usual, just
$$||f||_{L^p(T)} = \left(\int_T |f(t)|^p\,dt\right)^{1/p}.$$
We can extend this scale of norms to $p = \infty$ by taking
$$||f||_{L^\infty} = \sup_{t \in T} |f(t)|.$$
- *Sobolev-norms:* Let $1 \le p < \infty$. The L^p-Sobolev norm is
$$||f||_{W_p^m(T)} = ||f||_{L^p} + ||f^{(m)}||_{L^p}.$$
- *Hölder classes:* Let $0 < \alpha < 1$. The Hölder class $C^\alpha(T)$ is the collection of continuous functions f on the domain T with $|f(t) - f(t')| \le C|t - t'|^\alpha$ and $||f||_{L^\infty} < C$, for some $C > 0$; the smallest such C is the norm. Let $m < \alpha < m + 1$, for integer $m \ge 1$; the Hölder class $C^\alpha(T)$ is the collection of continuous functions f on the domain T with
$$|f^{(m)}(t) - f^{(m)}(t')| \le C|t - t'|^\delta$$
for $\delta = \alpha - m$.
- *Bump Algebra:* Suppose that f is a continuous function on the line $T = (-\infty, \infty)$. Let $g(t) = e^{-t^2}$ be a Gaussian normalized to height one rather than area one. Suppose that f can be represented as $\Sigma_i a_i g((t - t_i)/s_i)$ for a countable sequence of triples (a_i, t_i, s_i) with $s_i > 0$, $t_i \in T$, and $\Sigma_i |a_i| = C < \infty$. Then f is said to belong to the Bump Algebra, and its Bump norm $|f|_B$ is the smallest value of C for which such a decomposition exists. Evidently, a function in the Bump Algebra is a superposition of Gaussians, with various polarities, locations, and scales.
- *Bounded Variation:* Suppose that f is a function on the interval $T = [0, 1]$ that is integrable and such that the increment obeys
$$||f(\cdot + h) - f(\cdot)||_{L^1[0,1-h)} \le C|h|$$
for $0 < h < 1$. The BV seminorm of f is the smallest C for which this is true.

In each case, the norm equivalence problem is: *Find an orthobasis (ϕ_k) and a discrete norm $||\theta||_{\boldsymbol{f}}$ so that the F norm $||f||_F$ is equivalent to the discrete norm $||\theta(f)||_{\boldsymbol{f}}$.* In-depth discussions of these spaces and their norm-equivalence problems can be found in [103], [89], [90], [96], [43], and [74]. In some cases, as we explain in Section XII below, the norm equivalence has been solved; in other cases, it has been proven that there can never be a norm equivalence, but a closely related space has been discovered for which a norm equivalence is available.

In these five problems, Fourier analysis does not work; that is, one cannot find a "simple" and "natural" norm on the Fourier coefficients which provides an equivalent norm

to the considered continuous norm. It is also true that tight equivalence results, with $A = B = 1$, seem out of reach in these cases.

The key point in seeking a norm equivalence is that the discrete norm must be "simple" and "natural." By this we really mean that the discrete norm *should depend only on the size of the coefficients* and not on the signs or phases of the coefficients. We will say that the discrete norm is *unconditional* if it obeys the relationship

$$||\theta'||_{\boldsymbol{f}} \le ||\theta||_{\boldsymbol{f}}$$

whenever $\theta'_k = s_k\theta_k$ with s_k any sequence of weights $|s_k| \le 1$. The idea is that "shrinking the coefficients in size" should "shrink" the norm.

A norm equivalence result therefore requires discovering both a representing system (ϕ_k) and a special norm on a sequence space, one which is equivalent to the considered continuum norm and also has the unconditionality property. For future reference, we call a basis yielding a norm equivalence between a norm on function space and such an unconditional norm on sequence space an *unconditional basis*. It has the property that for any object in a function class F, and any set of coefficients θ'_k obeying

$$|\theta'_k| \le |\theta_k(f)|, \qquad \forall k$$

the newly constructed object

$$f' = \sum_k \theta'_k \phi_k$$

belongs to F as well.

There is a famous result dramatizing the fact that the Fourier basis is not an unconditional basis for classes of continuous functions, due to DeLeeuw, Kahane, and Katznelson [25]. Their results allows us to construct pairs of functions: one, g, say, which is uniformly continuous on the circle; and the other, h, say, very wild, having square integrable singularities on a dense subset of the circle. The respective Fourier coefficients obey

$$|\theta_k(g)| > |\theta_k(h)| \qquad \forall k.$$

In short, the ugly and bizarre object h has the *smaller* Fourier coefficients. Another way of putting this is that an extremely delicate pattern in the phases of the coefficients, rather than the size of the coefficients, control the regularity of the function. The fact that special "conditions" on the coefficients, unrelated to their size, are needed to impose regularity may help to explain the term "unconditional" and the preference for unconditional structure.

XII. Harmonic Analysis and Norm Equivalence

We now describe a variety of tools that were developed in harmonic analysis over the years in order to understand norm equivalence problems, and some of the norm equivalence problems that were solved.

A. Warmup: The Sampling Theorem

A standard procedure by which interesting orthobases have been constructed by harmonic analysts is to first develop a kind of "overcomplete" continuous representation, and later develop a discretized variant based on a geometric model.

An elementary example of this procedure will be familiar to readers in the information theory community, as Shannon's Sampling Theorem [84] in signal analysis. Let $\hat{f}(\omega)$ be an L^2 function supported in a finite frequency interval $[-\pi\Omega, \pi\Omega]$; and let $f(t) = (1/2\pi) \int \hat{f}(\omega) \exp\{i\omega t\}\, d\omega$ be the time-domain representation. This representation is an L^2-isometry, so that

$$\frac{1}{2\pi} \int_{-\pi\Omega}^{\pi\Omega} |\hat{f}(\omega)|^2 = \int_T |f(t)|^2\, dt.$$

The size of the function on the frequency side and on the time side are the same, up to normalization. By our assumptions, the time-domain representation $f(t)$ is a bandlimited function, which is very smooth, and so the representation of f in the time domain is very redundant. A nonredundant representation is obtained by sampling, and retaining only the $f(k/\Omega)$. The mathematical expression of the perfect nonredundancy of this representation is the fact that we have the norm equivalence

$$||f||_{L^2(\boldsymbol{R})} = \left(\frac{1}{\Omega} \sum_k |f(k/\Omega)|^2 \right)^{1/2} \tag{12.1}$$

and that there is an orthobasis of sampling functions $\varphi_k(t) = \Omega^{1/2} \operatorname{sinc}(\Omega t - k)$ so that $f(k/\Omega) = \langle \varphi_k, f \rangle \cdot \Omega^{1/2}$ and

$$f(t) = \sum_k f(k/\Omega)\varphi_k(t).$$

This time-domain representation has the following geometric interpretation: there is a sequence of disjoint "cells" of length $1/\Omega$, indexed by k; the samples summarize the behavior in those cells; and the sampling functions provide the details of that behavior. While this circle of ideas was known to Shannon and was very influential for signal analysis, we should point out that harmonic analysts developed ideas such as this somewhat earlier, under more general conditions, and based on an exploration of a geometric model explaining the phenomenon. For example, work of Paley and Wiener in the early 1930's and of Plancherel and Polya in the mid 1930's concerned norm equivalence in the more general case when the points of sampling were not equispaced, and obtained methods giving equivalence for all L^p norms, $p > 0$

$$||f||_{L^p(\boldsymbol{R})} \asymp \left(\frac{1}{\Omega} \sum_k |f(t_k)|^p \right)^{1/p} \tag{12.2}$$

provided the points t_k are approximately equispaced at density $1/\Omega$.

The results that the harmonic analysts obtained can be interpreted as saying that the geometric model of the sampling theorem has a very wide range of validity. For this geometric model, we define a collection of "cells" I_k, namely, intervals of length $1/\Omega$ centered at the sampling points t_k, and construct

the piecewise-constant object $\tilde{f} = \Sigma_k f(t_k)1_{I_k}(t)$. The norm equivalence (12.2) says, in fact, that

$$||f||_{L^p(\boldsymbol{R})} \asymp ||\tilde{f}||_{L^p(\boldsymbol{R})} \tag{12.3}$$

for a wide range of p.

This pattern—continuous representation, discretization, geometric model—has been of tremendous significance in harmonic analysis.

B. Continuous Time-Scale Representations

At the beginning of this century, a number of interesting relationships between Fourier series and harmonic function theory were discovered, which showed that valuable information about a function f defined on the circle T or the line $\mathbb{R}$ can be garnered from its harmonic extension into the interior of the circle, respectively the upper half plane; in other words, by viewing f as the boundary values of a harmonic function U. This theory also benefits from an interplay with complex analysis, since a harmonic function U is the real part of an analytic function $F = U + i\tilde{U}$. The imaginary part $\tilde{U}$ is called the conjugate function of U. The harmonic function U and the associated analytic function F give much important information about f. We want to point out how capturing this information on f through the functions $U(\cdot, y)$ ultimately leads to favorable decompositions of f into fundamental building blocks called "atoms." These decompositions can be viewed as a precursor to wavelet decompositions. For expository reasons we streamline the story and follow [89, Chs. III, IV], [43, Ch. 1], and [45]. Let f be any "reasonable" function on the line. It has a harmonic extension U into the upper half plane given by the Poisson integral

$$U(t,y) = \int_{\mathbf{R}} P_y(u)f(t-u)\,du, \qquad y>0, \tag{12.4}$$

where $P_y(t) = \pi^{-1}y/(y^2+t^2)$ is the Poisson kernel. This associates to a function f of one variable the harmonic function U of two variables, where the argument t again ranges over the line and y ranges over the positive reals. The physical interpretation of this integral is that f is the boundary value of U and $U(t,y)$ is what can be sensed at some "depth" y. Each of the functions $U(\cdot, y)$ is infinitely differentiable. Whenever $f \in L^p(\mathbb{R})$, $1 \le p \le \infty$, the equal-depth sections $U(\cdot, y)$ converge to f as $y \to 0$ and, therefore, the norms converge as well: $||U(\cdot,y)||_{L^p} \to ||f||_{L^p}$. We shall see next another way to capture $||f||_{L^p}$ through the function $U(\cdot, y)$. Hardy in 1914 developed an identity in the closely related setting where one has a function defined on the unit circle, and one uses the harmonic extension into the interior of the unit disk [53]; he noticed a way to recover the norm of the boundary values from the norm of the values on the interior of the disk. By conformal mapping and some simple identities based on Green's theorem, this is recognized today as equivalent to the statement that in the setting (12.4) the L^2 norm of the "boundary values" $f(t)$ can be recovered from the behavior of the whole function $U(t,y)$

$$\int |f(t)|^2\,dt = 8\pi \iint \left|\frac{\partial}{\partial y}U(t,y)\right|^2 y\,dy\,dt. \tag{12.5}$$

Defining $V(t,y) = 2\sqrt{2\pi}\cdot y\cdot(\partial/\partial y)U(t,y)$, this formula says

$$\int_{\mathbb{R}} |f(t)|^2\,dt = \int_{\mathbb{R}}\int_0^\infty |V(t,y)|^2 \frac{dy}{y}\,dt. \tag{12.6}$$

Now the object $V(t,y)$ is itself an integral transform

$$V(t,y) = \int Q_y(u)f(t-u)\,du$$

where the kernel $Q_y(u) = 2\sqrt{2\pi}\cdot y\cdot(\partial/\partial y)P_y(u)$ results from differentiating the Poisson kernel. This gives a method of assigning, to a function of a single real variable, a function of two variables, in an L^2 norm-preserving way. In addition, this association is invertible since, formally,

$$\begin{aligned} f(t) &= \lim_{y'\to 0} U(t,y') = \lim_{y'\to 0} \frac{1}{2\sqrt{2\pi}} \int_{y'}^\infty V(t,y)\frac{dy}{y} \\ &= \frac{1}{2\sqrt{2\pi}} \int_0^\infty V(t,y)\frac{dy}{y} \end{aligned}$$

when f is a "nice" function. In the 1930's, Littlewood and Paley, working again in the closely related setting of functions defined on the circle, found a way to obtain information on the L^p norms of a function defined on the circle, from an appropriate analysis of its values inside the circle. This is now understood as implying that V contains not just information about the L^2 norm as in (12.6), but also on L^p norms, $p \ne 2$. Defining $g_2(t) = (\int_0^\infty |V(t,y)|^2(dy/y)^{1/2}$, Littlewood–Paley theory in its modern formulation says that

$$||f||_{L^p} \asymp ||g_2||_{L^p} \tag{12.7}$$

for $1<p<\infty$. The equivalence (12.7) breaks down when $p \to 1$. (We shall not discuss the other problem point $p = \infty$ here since these L^∞-spaces are not separable and therefore do not have the series representations we seek.) To understand the reason for this breakdown one needs to examine the conjugate function $\tilde{U}$ of U mentioned above. It enjoys the same properties of U provided $1<p<\infty$. It has boundary values $\tilde{f}$ and the function $\tilde{f}$ (called the conjugate function of f) is also in $L^p(\mathbb{R})$. But the story takes a turn for the worse when $p = 1$: for a function $f \in L^1(\mathbb{R})$, its conjugate function $\tilde{f}$ need not be in $L^1(\mathbb{R})$. The theory of the real Hardy spaces H^p is a way to repair the situation and better understand the norm equivalences (12.7). A function f is said to be in real H^p, $1 \le p<\infty$, if and only if both f and its conjugate function $\tilde{f}$ are in L^p; the norm of f in H^p is the sum of the L^p norms of f and $\tilde{f}$. Replacing $||f||_{L^p}$ by $||f||_{H^p}$ on the left side of (12.7), we obtain equivalences with absolute constants that hold even for $p = 1$. In summary, the spaces H^p are a natural replacement for L^p when discussing representations and norm equivalences.

In modern parlance, the object V would be called an instance of *continuous wavelet transform*. This terminology was used for the first time in the 1980's by Grossmann and Morlet [52], who proposed the study of the integral transform

$$Wf(a,b) = \int_{-\infty}^\infty f(t)\overline{\psi}_{a,b}(t)\,dt \tag{12.8}$$

where

$$\psi_{a,b}(t) = |a|^{-1/2}\psi((t-b)/a)$$

and ψ is a so-called "wavelet," which must be chosen to obey an *admissibility condition*; for convenience, we shall require here a special form of this condition

$$2\pi = \int_0^\infty |\hat{\psi}(\xi)|^2|\xi|^{-1}\,d\xi = \int_{-\infty}^0 |\hat{\psi}(\xi)|^2|\xi|^{-1}\,d\xi. \quad (12.9)$$

Here b is a location parameter and a is a scale parameter, and the transform maps f into a time-scale domain. (Under a different terminology, a transform and admissibility condition of the same type can also be found in [2].) The wavelet transform with respect to an admissible wavelet is invertible

$$f(t) = \iint Wf(a,b)\psi_{a,b}(t)\frac{da}{a^2}\,db \quad (12.10)$$

and also an isometry

$$||f||_{L^2}^2 = \iint |Wf(a,b)|^2\frac{da}{a^2}\,db. \quad (12.11)$$

One sees by simple comparison of terms that the choice $\psi(t) = Q_1(t)$ yields $V(t,y) = W(a,b)$ under the calibration $a = y$ and $t = b$. We thus gain a new interpretation of V: the function Q_y has "scale" y and so we can interpret $V(t,y)$ as providing an association of the object f with a certain "time-scale" portrait.

The continuous wavelet transform is highly redundant, as the particular instance V shows: it is harmonic in the upper half-plane. Insights into the redundancy of the continuous wavelet transform are provided (for example) in [23]. Suppose we associate to the point (b,a) the rectangle

$$\rho(a,b) = [b-a, b+a] \times [a/2, 2a]$$

then the information "near (b,a)" in the wavelet transform is weakly related to information near (b',a') if the corresponding rectangles are well-separated.

C. Atomic Decomposition

As pointed out earlier, it is natural to replace the study of the spaces L^p with that of the spaces H^p; in particular, we avoid certain unsatisfactory aspects of L^1, which does not have any unconditional basis, and which behaves quite unlike its logically neighboring spaces L^p for $p > 1$. The norm equivalence (12.7), which does not work at $p = 1$, is then replaced by

$$||f||_{H^p} \asymp ||g_2||_{L^p}$$

valid for all $1 \le p < \infty$.

In the late 1960's and early 1970's, one of the most successful areas of research in harmonic analysis concerned H^1 and associated spaces. At that time, the concept of "atomic decomposition" arose, the key point was the discovery by Fefferman that one can characterize membership in the space H^1 precisely by the properties of its atomic decomposition into atoms obeying various size, oscillation, and support constraints. Since then "atomic decomposition" has come to mean a decomposition of a function f into a discrete superposition of nonrigidly specified pieces, where the pieces obey various analytic constraints (size, support, smoothness, moments) determined by a space of interest, with the size properties of those pieces providing the characterization of the norm of the function [16], [29], [42], [43].

The continuous wavelet transform gives a natural tool to build atomic decompositions for various spaces. Actually, the tool predates the wavelet transform, since already in the 1960's Calderón [9] established a general decomposition, a "resolution of identity operator" which in wavelet terms can be written

$$\int_{-\infty}^\infty \int_0^\infty \langle \cdot, \psi_{a,b}\rangle \psi_{a,b}\frac{da}{a^2}\,db = Id.$$

The goal is not just to write the identity operator in a more complicated form; by decomposing the integration domain using a partition into disjoint sets, one obtains a family of nontrivial operators, corresponding to different time-scale regions, which sum to the identity operator.

Let now I denote a *dyadic interval*, i.e., an interval of the form $I = I_{j,k} = [k/2^j, (k+1)/2^j)$ with j and k integers, and let $R(I)$ denote a time-scale (b,a)-rectangle sitting "above" I

$$R(I) = I \times (2^{-j-1}, 2^{-j}].$$

The collection $\mathcal{I}$ of all dyadic intervals I is obtained as the scale index j runs through the integers (both positive and negative) and the position index k runs through the integers as well. The corresponding collection $\mathcal{R}$ of all rectangles $R(I)$ forms a disjoint cover of the whole (a,b) plane; compare Fig. 4. If we now take the Calderón reproducing formula and partition the range of integration using this family of rectangles we have formally that $Id = \Sigma_I A_I$, where

$$A_I f = \int\int_{R(I)} \langle f, \psi_{a,b}\rangle \psi_{a,b}\frac{da}{a^2}\,db.$$

Here A_I is an operator formally associating to f that "piece" coming from time-scale region $R(I)$. In fact, it makes sense to call $A_I f$ a time-scale atom [43]. The region $R(I)$ of the wavelet transform, owing to the redundancy properties, constitutes in some sense a minimal coherent piece of the wavelet transform. The corresponding atom summarizes the contributions of this coherent region to the reconstruction, and has properties one would consider natural for such a summary. If ψ is supported in $[-1,1]$, then $A_I f$ will be supported in $3 \cdot I$, the interval with same center as I but three times its width. Also, if ψ is smooth and admissible then $A_I f$ will be smooth, oscillating only as much as required to be supported in $3 \cdot I$. For example, if f is m-times differentiable, and the wavelet ψ is chosen appropriately,

$$\left|\frac{\partial^m}{\partial t^m}(A_I f)(t)\right| \le C(\psi) \cdot |I|^{-m+1/2} \cdot ||f||_{C^m(5 \cdot I)} \quad (12.12)$$

and we cannot expect a better dependence of the properties of an atom on f in general. So the formula $f = \Sigma_I A_I f$ decomposes f into a countable sequence of time-scale atoms.

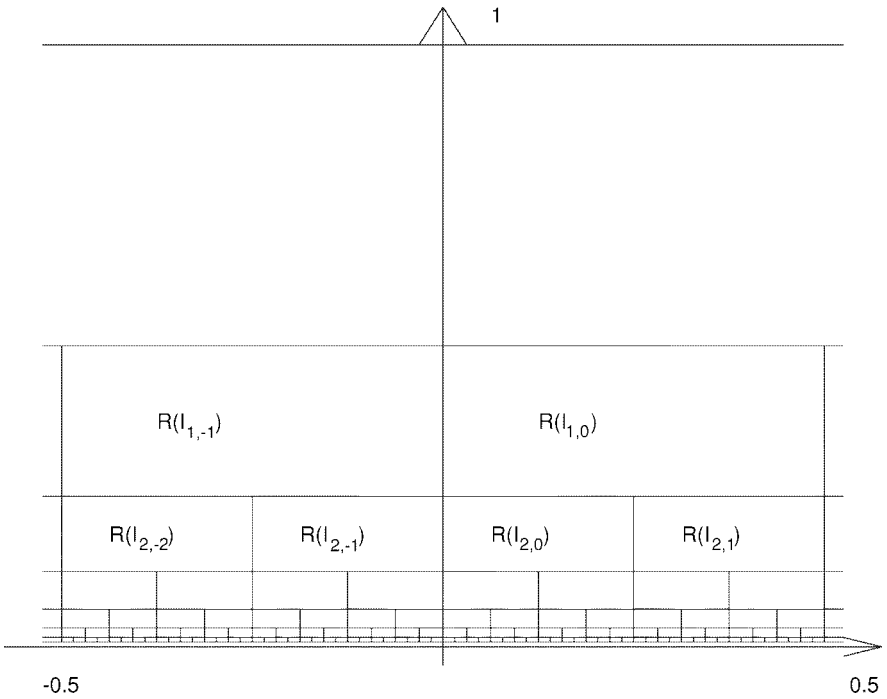

Fig. 4. A tiling of the time-scale plane by rectangles.

This analysis into atoms is informative about the properties of f; for example, suppose that f "is built from many large atoms at fine scales" then from (12.12) f cannot be very smooth. As shown in [43], one can express a variety of function space norms in terms of the size properties of the atoms; or, more properly, of the continuous wavelet transform associated to the "cells" $R(I)$. For example, if our measure of the "size" of the atom $A_I f$ is the "energy" of the "associated time-scale cell"

$$s_I(f) = \left(\int\int_{R(I)} |Wf(a,b)|^2 (da/a^2)\, db \right)^{1/2}$$

then $\Sigma_I s_I^2 = ||f||_{L^2}^2$.

D. Unconditional Basis

By the end of the 1970's, atomic decomposition methods gave a satisfactory understanding of H^p spaces; but in the early 1980's, the direction of research turned toward establishing a basis for these spaces. B. Maurey had shown by abstract analysis that an unconditional basis of H^1 must exist, and Carleson had shown a way to construct such a basis by. in a sense. repairing the lack of smoothness of the Haar basis.

Let $h(t) = 1_{[1/2,1)}(t) - 1_{[0,1/2)}(t)$ denote the Haar function; associate a scaled and translated version of h to each dyadic interval $I = 2^{-j}[k, k, +1]$ according to $h_I(t) = 2^{j/2} h(2^j t - k)$. This collection of functions makes a complete orthonormal system for $L^2(R)$. In particular, $f = \Sigma_I \langle f, h_I \rangle h_I$. This is similar in some ways to the atomic decomposition idea of the last section—it is indexed by dyadic intervals and associates a decomposition to a family of dyadic intervals. However, it is better in other ways, since the h_I are fixed functions rather than atoms. Unfortunately, while it gives an unconditional basis for L^2, it does not give an unconditional basis for most of the spaces where Littlewood–Paley theory applies; the discontinuities in the h_I are problematic.

Carleson's attempt at 'repairing the Haar basis' stimulated Stromberg (1982), who succeeded in constructing an orthogonal unconditional basis for H^p spaces with $0 < p \leq 1$. In effect, he showed how, for any $p \leq 1$, to construct a function ψ so that the functions $\psi_I(t) = 2^{j/2} \psi(2^j t - k)$ were orthogonal and constituted an unconditional basis for H^p. Today we would call such a ψ a wavelet; in fact, Stromberg's ψ is a spline wavelet—a piecewise polynomial—and (ψ_I) constituted the first orthonormal wavelet basis with smooth elements. In [92], Stromberg testifies that at the time that he developed this basis he was not aware of, or interested in, applications outside of harmonic analysis. Stromberg's work was published in a conference proceedings volume and was not widely noticed at the time.

In the mid-1980's, Yves Meyer and collaborators became very interested in the continuous wavelet transform as developed by Grossmann and Morlet. They first built frame expansions, which are more discrete than the continuous wavelet transform and more rigid than the atomic decomposition. Then, Meyer developed a bandlimited function ψ—the Meyer wavelet—which generated an orthonormal basis for L^2 having elements ψ_I which were infinitely differentiable and decayed rapidly at $\pm\infty$. Lemarié and Meyer [64] then showed that this offered an unconditional basis for a very wide range of spaces: all the L^p-Sobolev, $1 < p < \infty$, of all orders $m = 0, 1, 2, \cdots$; all the Hölder H^1, and more generally, all Besov spaces. Frazier and Jawerth have shown that the Meyer basis offers unconditional bases for all the spaces in the Triebel scale; see [43].

E. Equivalent Norms

The solution of these norm-equivalence problems required the introduction of two new families of norms on sequence space; the general picture is due to Frazier and Jawerth [42], [43].

The first family is the (homogeneous) *Besov sequence norms*. With $\theta_{j,k}$ a shorthand for the coefficient θ_I arising from the dyadic interval $I = I_{j,k} = [k/2^j, (k+1)/2^j)$,

$$||\theta||_{\dot{\boldsymbol{b}}^{\alpha}_{p,q}} = \left(\sum_j 2^{j(\alpha+1/2-1/p)q} \left(\sum_k |\theta_{j,k}|^p \right)^{q/p} \right)^{1/q},$$

with an obvious modification if either p or $q = \infty$. These norms first summarize the sizes of the coefficients at all positions k with a single scale j using an ℓ^p norm and then summarize across scales, with an exponential weighting.

The second family involves the (homogeneous) *Triebel sequence norms*. With χ_I the indicator of $I = I_{j,k}$,

$$||\theta||_{\dot{\boldsymbol{f}}^{\alpha}_{pq}} = \left\| \left(\sum_I (2^{j(\alpha+1/2)} \chi_I(t) |\theta_I|)^q \right)^{1/q} \right\|_{L^p},$$

with an obvious modification if $p = \infty$ and a special modification if $q = \infty$ (which we do not describe here; this has to do with the space BMO). These norms first summarize the sizes of the coefficients across scales, and then summarize across positions.

The reader should note that the sequence space norm expressions have the unconditionality property: they only involve the sizes of the coefficients. If one shrinks the coefficients in such a way that they become term-wise smaller in absolute values, then these norms must decrease.

We give some brief examples of norm equivalences using these families. Recall the list of norm equivalence problems listed in Section XI. Our first two examples use the Besov sequence norms.

- (Homogeneous) Hölder Space $\dot{C}^{\alpha}(\boldsymbol{R})$: For $0 < \alpha < 1$, the norm $||f||_{\dot{C}^{\alpha}}$ is the smallest constant C so that
$$|f(t) - f(t')| \le C|t - t'|^{\alpha}.$$
To get an equivalent norm, use an orthobasis built from (say) Meyer wavelets, and measure the norm of the wavelet coefficients by the Besov norm with α, $p = q = \infty$. This reduces to a very simple expression, namely,
$$||\theta||_{\dot{\boldsymbol{b}}^{\alpha}_{\infty,\infty}} = \sup_I |\theta_I| 2^{j(\alpha+1/2)}. \tag{12.13}$$
In short, membership in $\dot{C}^{\alpha}$ requires that the wavelet coefficients *all* decay like an appropriate power of the associated scale: $|\theta_I| \le C' \cdot 2^{-j(\alpha+1/2)}$.
- *Bump Algebra:* Use an orthobasis built from the Meyer wavelets, and measure the norm of the wavelet coefficients by the Besov norm with $\alpha = 1, p = q = 1$. This reduces to a very simple expression, namely,
$$||\theta||_{\dot{\boldsymbol{b}}^{1}_{1,1}} = \sum_I |\theta_I| 2^{j/2}. \tag{12.14}$$
Membership in the Bump Algebra thus requires that the sum of the wavelet coefficients decay like an appropriate power of scale. Note, however, that some coefficients could be large, at the expense of others being correspondingly small in order to preserve the size of the sum.

We can also give examples using the Triebel sequence norms.

- L^p: For $1 < p < \infty$, use an orthobasis built from, e.g., the Meyer wavelets, and measure the norm of the wavelet coefficients by the Triebel sequence norm with $\alpha = 0$, $q = 2$, and p precisely the same as the p in L^p.
- *L^p-Sobolev spaces* W^m_p: For $1 < p < \infty$, use an orthobasis built from, e.g., the Meyer wavelets, and measure the norm of the wavelet coefficients by a superposition of two Triebel sequence norms, one with $\alpha = 0$, $q = 2$, and p precisely the same as the p in L^p; the other with $\alpha = m$, $q = 2$, and p precisely the same as the p in L^p.

F. Norm Equivalence as a Sampling Theorem

The Besov and Triebel sequence norms have an initially opaque appearance. A solid understanding of their structure comes from a view of norm equivalence as establishing a sampling theorem for the upper half-plane, in a manner reminiscent of the Shannon sampling theorem and its elaborations (12.1) and (12.2).

Recall the Littlewood–Paley theory of the upper half plane of Section XII-A and the use of dyadic rectangles $R(I)$ built "above" dyadic intervals from Section XII-B. Partition the upper half-plane according to the family $R(I)$ of rectangles. Given a collection of wavelet coefficients $\theta = (\theta_I)$, assign to rectangle $R(I)$ the value $|\theta_I|$. One obtains in this way a pseudo-V-function

$$\tilde{V}(t,y) = \sum_I |\theta_I| 1_{R(I)}(t,y).$$

This is a kind of caricature of the Poisson integral V. Using this function as if it were a true Poisson integral suggests to calculate, for $q < \infty$,

$$\tilde{g}_q(t) = \left(\int_0^{\infty} |\tilde{V}(t,y)|^q \frac{dy}{y} \right)^{1/q}.$$

As it turns out, the Triebel sequence norm is *precisely* a simple continuum norm of the object g_q

$$||\theta||_{\dot{\boldsymbol{f}}^{0}_{p,q}} = ||\tilde{g}_q||_{L^p}.$$

In short, the Triebel sequence norm expresses the geometric analogy that the piecewise-constant object $\tilde{V}$ may be treated as if it were a Poisson integral. Why is this reasonable?

Observe that the wavelet coefficient θ_I is *precisely* a sample of the continuous wavelet transform with respect to the wavelet ψ generating the orthobasis ψ_I

$$\theta_I = \langle f, \psi_I \rangle = (W_{\psi} f)(2^{-j}, k2^{-j}).$$

The sampling point is $(a_j, b_{j,k})$, where $a_j = 2^{-j}, b_{j,k} = k/2^j$; these are the coordinates of the lower left corner of the

rectangle $R(I_{j,k})$. In the $\alpha = 0$ case, we have

$$\tilde{V}(t,y) = \sum_{j,k} |W_\psi f(2^{-j}, k2^{-j})| 1_{R(I_{j,k})}(t,y).$$

That is, $\tilde{V}$ is a pseudo-continuous wavelet transform, gotten by replacing the true continuous wavelet transform on each cell by a cell-wise constant function with the same value in the lower left corner of each cell.

The equivalence of the true L^p norm with the Triebel sequence norm of the wavelet coefficients expresses the fact that the piecewise-constant function $\tilde{V}$, built from time-scale samples of Wf has a norm—defined on the whole time-scale plane—which is equivalent to Wf. Indeed, define a norm on the time-scale plane by

$$||U||_{T_{p,q}} = \left(\int_{-\infty}^{\infty} \left(\int_0^\infty |U(t,y)|^q \frac{dy}{y} \right)^{p/q} dt \right)^{1/p}$$

summarizing first across scales and then across positions. We have the identity $||\tilde{g}_q||_{L_p} = ||\tilde{V}||_{T_{p,q}}$. The equivalence of norms $||f||_{L^p} \asymp ||\theta(f)||_{\mathbf{f}^0_{p,2}}$ can be broken into the following stages:

$$||f||_{L^p} \asymp ||V||_{T_{p,2}}$$

which follows from Littlewood–Paley theory,

$$||V||_{T_{p,2}} \asymp ||Wf||_{T_{p,2}}$$

which says that the "Poisson wavelet" Q_1, and some other nice wavelet ψ obtain equivalent information, and finally

$$||Wf||_{T_{p,2}} \asymp ||\tilde{V}||_{T_{p,2}}.$$

This is a sampling theorem for the upper half-plane, showing that an object Wf and its piecewise-constant approximation $\tilde{V}$ have equivalent norms. It is exactly analogous to the equivalence (12.3) that we discussed in the context of the Shannon sampling theorem. Similar interpretations can be given for the Besov sequence norm as *precisely* a simple continuum norm of the object $\tilde{V}$. In the case $p, q < \infty$

$$||\theta||_{\mathbf{b}^0_{p,q}} = \left(\int_0^\infty \left(\int_{-\infty}^{\infty} |\tilde{V}(t,y)|^p \, dt \right)^{q/p} \frac{dy}{y} \right)^{1/q}.$$

The difference is that the Besov norm involves first a summarization in position t and then a summarization in scale y; this is the opposite order from the Triebel case.

XIII. Norm Equivalence and Axes of Orthosymmetry

There is a striking geometric significance to the unconditional basis property.

Consider a classical example using the exact norm equivalence properties (11.3) and (11.4) from Fourier analysis. Suppose we consider the class $W^m_{2,0}(\gamma)$ consisting of all functions f obeying

$$||f||_{W^m_{2,0}} \le \gamma$$

with

$$||f||^2_{W^m_{2,0}} = ||f||^2_{L^2} + ||f^{(m)}||^2_{L^2}.$$

This is a body in infinite-dimensional space; defined as it is by quadratic constraints, we call it an ellipsoid. Owing to the exact norm equivalence properties (11.3), (11.4), the axes of symmetry of the ellipsoid are precisely the sinusoids (ϕ_k).

Something similar occurs in the nonclassical cases. Consider, for example, the Hölder norm equivalence (12.13). This says that, up to an equivalent re-norming, the Hölder class is a kind of hyperrectangle in infinite-dimensional space. This hyperrectangle has axes of symmetry; the directions of these axes are given by the members of the wavelet basis.

Consider now the Bump Algebra norm equivalence (12.14). This says that, up to an equivalent re-norming, the Bump Algebra is a kind of octahedron in infinite-dimensional space. This octahedron has axes of symmetry, and these are again given by the members of the wavelet basis.

So the *unconditional basis property* means that the *basis functions serve as axes of symmetry* for the corresponding function ball. This is analogous to the existence of axes of symmetry for an ellipsoid, but is more general: it applies in the case of function balls which are not ellipsoidal, i.e., not defined by quadratic constraints.

There is another way to put this that might also be useful. The axes of orthosymmetry of an ellipsoid can be derived as eigenfunctions of the quadratic form defining the ellipsoid. The axes of orthosymmetry solve the problem of "rotating the space" into a frame where the quadratic form becomes diagonal. In the more general setting, where the norm balls are not ellipsoidal, we can say that an unconditional basis solves the problem of "rotating the space" into a frame where the norm, although not involving a quadratic functional, is "diagonalized."

XIV. Best Orthobasis for Nonlinear Approximation

The unconditional basis property has important implications for schemes of nonlinear approximation which use the best n-terms in an orthonormal basis. In effect, the unconditional basis of a class F will be, in a certain asymptotic sense, the best orthonormal basis for n-term approximation of members of the associated function ball $\mathcal{F}$. We highlight these results and refer the reader to [26] and [30] for more details on nonlinear approximation.

A. n-Term Approximations: Linear and Nonlinear

Suppose one is equipped with an orthobasis (ϕ_k), and that one wishes to approximate a function f using n-term expansions

$$f \approx P_n(f; (\phi_k), (k_i)) \equiv \sum_{i=1}^{n} a_i \phi_{k_i}.$$

If the k_i are fixed—for example as the first n-basis elements in the assumed ordering—this is a problem of linear approximation, which can be solved (owing to the assumed orthogonality of the ϕ_k) by taking $a_i = \langle f, \phi_{k_i} \rangle$. Supposing that the $(k_i: i = 1, 2, \cdots)$ is an enumeration of the integers, the approximation error in such an orthogonal system is $\Sigma_{i>n} a_i^2$, which means that the error is small if the important coefficients occur in the leading n-terms rather than in tail of the sequence.

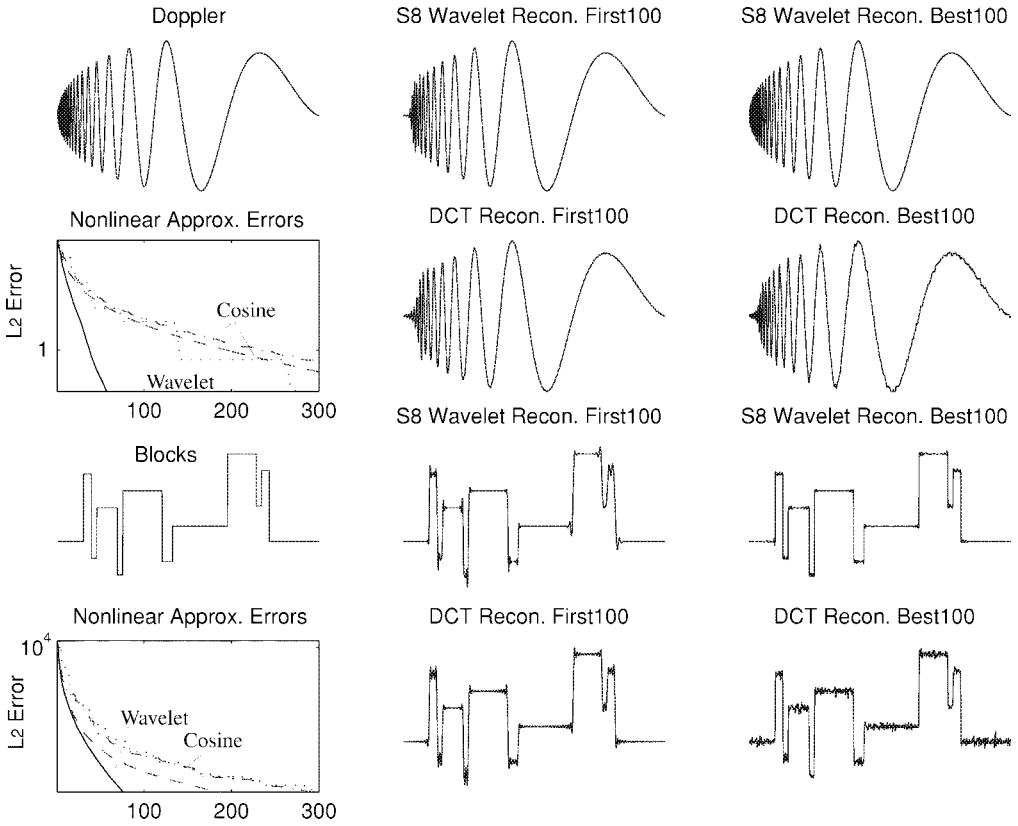

Fig. 5. Two objects, and performance of best-n and first-n approximations in both Fourier and wavelet bases, and linear and nonlinear approximation numbers (14.2).

Such linear schemes of approximation seem very natural when the orthobasis is either the classical Fourier basis or else a classical orthogonal polynomial basis, and we think in terms of the "first n-terms" as the "n lowest frequencies." Indeed, several results on harmonic analysis of smooth functions promote the expectation that the coefficients $\langle f, \phi_k \rangle$ decay with increasing frequency index k. It therefore seems natural in such a setting to adopt once and for all the standard ordering $k_i = i$, to expect that the a_i decay with i, or in other words, that the important terms in the expansion occur at the early indices in the sequence. There are many examples in both practical and theoretical approximation theory showing the essential validity of this type of linear approximation.

In general, orthobases besides the classical Fourier/orthogonal polynomial sets (and close relatives) cannot be expected to display such a fixed one-dimensional ordering of coefficient amplitudes. For example, the wavelet basis has two indices, one for scale and one for position, and it can easily happen that the "important terms" in an expansion cluster at coefficients corresponding to the same position but at many different scales. This happens, for example, when we consider the coefficients of a function with punctuated smoothness, i.e., a function which is piecewise-smooth away from discontinuities. Since, in general, the position of such singularities varies from one function to the next, in such settings, one cannot expect the "n-most important coefficients" to occur in indices chosen in a fixed nonadaptive fashion. It makes more sense to consider approximation schemes using n-term approximations where the n-included terms are the *biggest-n-terms* rather then the *first-n-terms* in a fixed ordering; compare Fig. 5. Equivalently, we consider the ordering $k_i^* \equiv k_i^*(f)$ defined by coefficient amplitudes

$$|a_{k_1^*}^*| \geq |a_{k_2^*}| \geq \cdots \geq |a_{k_i^*}| \geq |a_{k_{i+1}^*}| \geq \cdots \tag{14.1}$$

and define

$$Q_n(f; (\phi_k)) = \sum_{i=1}^{n} a_i \phi_{k_i^*(f)}$$

where again $a_i = \langle f, \phi_{k_i^*(f)} \rangle$. This operator at first glance seems linear, because the functionals $a_i(f)$ derive from inner products of f with basis functions $\phi_{k_i^*}$; however, in fact it is nonlinear, because the functionals $k_i^*(f)$ depend on f.

This nonlinear approximation operator conforms in the best possible fashion with the idea that the n-most-important coefficients should appear in the approximation, while the less important coefficients should not. It is also quantitatively better than any fixed linear approximation operator built from the basis (ϕ_k)

$$\|f - Q_n(f; (\phi_k))\|_{L^2} = \min_{(k_i)} \|f - P_n(f; (\phi_k), (k_i))\|_{L^2}$$

because the square of the left-hand side is $\Sigma_{i>n} \langle f, \phi_{k_i^*(f)} \rangle^2$, which by the rearrangement relation (14.1) is not larger than any sum $\Sigma_{i>n} \langle f, \phi_{k_i} \rangle^2$.

B. Best Orthobasis for n-Term Approximation

In general, it is not possible to say much interesting about the error in n-term approximation for a fixed function f. We therefore consider the approximation for a function class $\mathcal{F}$ which is a ball $\{f\colon \|f\|_F \le \gamma\}$ of a normed (or quasinormed) linear space F of functions. Given such a class $\mathcal{F}$ and an orthogonal basis Φ, the error of best n-term approximation is defined by

$$d_n(\Phi, \mathcal{F}) = \sup_{f \in \mathcal{F}} \|f - Q_n(f; \Phi)\|_{L^2}. \tag{14.2}$$

We call the sequence $(d_n(\mathcal{F}))_{n=1}^{\infty}$ the Stechkin numbers of $\mathcal{F}$, honoring the work of Russian mathematician S. B. Stechkin, who worked with a similar quantity in the Fourier basis. The Stechkin numbers give a measure of the quality of representation of $\mathcal{F}$ from the basis Φ. If $d_n(\mathcal{F}, \Phi)$ is small, this means that every element of $\mathcal{F}$ is well-approximated by n-terms chosen from the basis, with the possibility that different sets of n-terms are needed for different $f \in \mathcal{F}$.

Different orthonormal bases can have very different approximation characteristics for a given function class. Suppose we consider the class $\mathrm{BV}(1)$ of all functions on $T = [0, 2\pi)$, with bounded variation ≤ 1 as in [31]. For the Fourier system, $d_n(\mathrm{BV}(1), \textsc{Fourier}) \asymp n^{-1/2}$, owing to the fact that BV contains objects with discontinuities, and sinusoids have difficulty representing such discontinuities, while for the Haar-wavelet system $d_n(\mathrm{BV}(1), \textsc{Haar}) \asymp n^{-1}$, intuitively because the operator Q_n in the Haar expansion can easily adapt to the presence of discontinuities by including terms in the expansion at fine scales in the vicinity of such singularities and using only terms at coarse scales far away from singularities.

Consider now the problem of finding a best orthonormal basis—i.e., of solving

$$d_n^*(\mathcal{F}) = \inf_{\Phi} d_n(\Phi, \mathcal{F}).$$

We call $d_n^*(\mathcal{F})$ the nonlinear orthowidth of $\mathcal{F}$; this is not one of the usual definitions of nonlinear widths (see [77]) but is well suited for our purposes. This width seeks a basis, in some sense, ideally adapted to $\mathcal{F}$, getting the best guarantee of performance in approximation of members of $\mathcal{F}$ by the use of n adaptively chosen terms. As it turns out, it seems out of reach to solve this problem exactly in many interesting cases; and any solution might be very complex, for example, depending delicately on the choice of n. However, it is possible, for balls $\mathcal{F}$ arising from classes F which have an unconditional orthobasis and sufficient additional structure, to obtain asymptotic results. Thus for function balls $W_2^m(\gamma)$ defined by quadratic constraints on the function and its mth derivative, we have

$$\begin{aligned} d_n^*(W_2^m(\gamma)) &\asymp d_n(\textsc{Fourier}, W_2^m(\gamma)) \\ &\asymp d_n(\textsc{Periodic Wavelets}, W_2^m(\gamma)) \asymp n^{-m}, \quad n \to \infty \end{aligned}$$

saying that either the standard Fourier basis, or else a smooth periodic wavelet basis, is a kind of best orthobasis for such a class. For the function balls $\mathrm{BV}(\gamma)$ consisting of functions on the interval with bounded variation $\le \gamma$

$$d_n^*(\mathrm{BV}(\gamma)) \asymp d_n(\textsc{Haar}, \mathrm{BV}(\gamma)) \asymp n^{-1}, \qquad n \to \infty$$

so that the Haar basis is a kind of best orthobasis for such a class. For comparison, a smooth wavelet basis is also a kind of best basis, $d_n(\textsc{Wavelets}, \mathrm{BV}(\gamma)) \asymp n^{-1}$, while a Fourier basis is not: $d_n(\textsc{Fourier}, \mathrm{BV}(\gamma)) \asymp n^{-1/2} \ne O(n^{-1})$. A summary is given in the table at the bottom of this page.

Results on rates of nonlinear approximation by wavelet series for an interesting range Besov classes were obtained in [29].

C. Geometry of Best Orthobasis

There is a nice geometric picture which underlies these results about best orthobases; essentially, *an orthobasis unconditional for a function class F is asymptotically a best orthobasis for that class*. Thus the extensive work of harmonic analysts to build unconditional bases—a demanding enterprise whose significance was largely closed to outsiders for many years—can be seen as an effort which, when successful, has as a byproduct the construction of best orthobases for nonlinear approximation.

To see this essential point requires one extra element. Let $0 < p < 2$, and note especially that we include the possibility that $0 < p < 1$. Let $|\theta|_{(k)}$ denote the kth element in the decreasing rearrangement of magnitudes of entries in θ, so that $|\theta|_{(1)} \ge |\theta|_{(2)} \ge \cdots$. The weak-$\ell^p$-norm is

$$\|\theta\|_{w\ell^p} = \sup_{k>0} k^{1/p} |\theta|_{(k)}.$$

This is really only a quasinorm; it does not in general obey the triangle inequality. This norm is of interest because of the way it expresses the rate of approximation of operators Q_n in a given basis. Indeed, if $p = 2/(2m+1)$ then

$$\|f - Q_n f\|_2 \le C_1(p) \|\theta\|_{w\ell^p} \cdot (n+1)^{-m}, \qquad n \ge 0.$$

Name	F	Best Basis	$d_n^*(\mathcal{F}) \asymp$	$\alpha_n^*(\mathcal{F})$
L^2-Sobolev	W_2^m	Fourier or Wavelet	n^{-m}	m
L^p-Sobolev	W_p^m	Wavelet	n^{-m}	m
Hölder	$\dot{C}^{\alpha}$	Wavelet	$n^{-\alpha}$	α
Bump Algebra	$\dot{B}_{1,1}^1$	Wavelet	n^{-1}	1
Bounded Variation	BV	Haar	n^{-1}	1
Segal Algebra	S	Wilson	$n^{-1/2}$	1/2

The result has a converse

$$\sup_{n\geq 0}(n+1)^m \cdot ||f - Q_n f||_2 \geq C_0(p)||\theta||_{w\ell^p}.$$

A given function f can be approximated by Q_n with error $\leq C_1(n+1)^{-m}$ for all n if $||\theta||_{w\ell^p} \leq 1$; it can only be approximated with error $\leq C_0(n+1)^{-m}$ if $||\theta||_{w\ell^p} \leq 1$. A given function ball $\mathcal{F}(\gamma)$ has functions f all of which can be approximated by $Q_n(f;(\phi_k))$ at error $\leq C\cdot(n+1)^{-m}$ for all n and for some C fixed independently of f if and only if for some C'

$$\sup\{||\theta(f)||_{w\ell^p}\colon f \in \mathcal{F}(\gamma)\} < C'.$$

Suppose we have a ball $\mathcal{F}$ arising from a class F with an unconditional basis (ϕ_k) and we consider using a possibly different orthobasis (φ_k) to do the nonlinear approximation: $Q_n(\cdot;(\varphi_k))$. The coefficient sequence $\omega = (\langle\varphi_k, f\rangle)$ obeys $\omega = U\theta$ where U is an orthogonal transformation of ℓ^2. Define then $\Theta = \{\theta(f)\colon f \in \mathcal{F}(\gamma)\}$. The rate of convergence of nonlinear approximation with respect to the new system is constrained by the relation

$$\begin{aligned}\sup\{||\omega(f)||_{w\ell^p}\colon f \in \mathcal{F}(\gamma)\} &= \sup\{||U\theta(f)||_{w\ell^p}\colon f \in \mathcal{F}(\gamma)\}\\ &\geq \sup\{||U\theta||_{w\ell^p}\colon \theta \in \Theta\}.\end{aligned}$$

By the unconditional basis property, there is an orthosymmetric set Θ^0 and constants A and B such that $A\cdot\Theta^0 \subset \Theta \subset B\cdot\Theta^0$. Hence, up to homothetic multiples, Θ is orthosymmetric.

Now the fact that Θ^0 is orthosymmetric means that it is in some sense "optimally positioned" about its axes; hence it is intuitive that rotating by U cannot improve its position. Thus [31]

$$\sup\{||U\theta||_{w\ell^p}\colon \theta \in \Theta^0\} \geq c(p)\sup\{||\theta||_{w\ell^p}\colon \theta \in \Theta^0\}.$$

We conclude that if (ϕ_k) is an orthogonal unconditional basis for F and if $d_n(\mathcal{F}(\gamma);(\phi_k)) \asymp n^{-m}$, then it is impossible that $d_n^*(\mathcal{F}(\gamma)) \leq C\cdot n^{-m'}$ for any $m' > m$. Indeed, if it were so, then we would have some basis (φ_k) achieving $d_n(\mathcal{F}(\gamma);(\varphi_k)) \leq C\cdot n^{-m'}$, which would imply that also the unconditional basis achieves $d_n(\mathcal{F}(\gamma);(\phi_k)) \leq C'\cdot n^{-m'}, m' > m$, contradicting the assumption that merely $d_n(\mathcal{F}(\gamma);(\phi_k)) \asymp n^{-m}$. In a sense, up to a constant factor improvement, the axes of symmetry make a best orthogonal basis.

XV. ϵ-Entropy of Functional Classes

We now show that nonlinear approximation in an orthogonal unconditional basis of a class F can be used, quite generally, to obtain asymptotically, as $\epsilon \to 0$, the optimal degree of data compression for a corresponding ball $\mathcal{F}$. This completes the development of the last few sections.

A. State of the Art

Consider the problem of determining the Kolmogorov ϵ-entropy for a function ball $\mathcal{F}$; this gives the minimal number of bits needed to describe an arbitrary member of $\mathcal{F}$ to within accuracy ϵ.

This is a hard problem, with very few sharp results. Underscoring the difficulty of obtaining results in this area is the commentary of V. M. Tikhomirov in Kolmogorov's *Selected Works* [94]

> The question of finding the exact value of the ϵ-entropy ... is a very difficult one .... Besides [one specific example] ... the author of this commentary knows no meaningful examples of infinite-dimensional compact sets for which the problem of ϵ-entropy ... is solved exactly.

This summarizes the state of work in 1992, more than 35 years after the initial concept was coined by Kolmogorov. In fact, outside of the case alluded to by Tikhomirov, of ϵ-entropy of Lipschitz functions measured in L^∞ metric, and the case which has emerged since then, of smooth functions in L^2 norm as mentioned above, there are no asymptotically exact results. The typical state of the art for research in this area is that within usual scales of function classes having finitely many derivatives, one gets order bounds: finite positive constants A_0 and A_1, and an exponent m depending on $\mathcal{F}$ but not on $\epsilon < \epsilon_0$, such that

$$A_0\epsilon^{-1/m} \leq H_\epsilon(\mathcal{F}) \leq A_1\epsilon^{-1/m}, \qquad \epsilon \to 0. \tag{15.1}$$

Such order bounds display some paradigmatic features of the state of the art of ϵ-entropy research. First, such a result *does* tie down the precise rate involved in the growth of the net (i.e., $H_\epsilon = O(\epsilon^{-1/m})$ as $\epsilon \to 0$). Second, it *does not* tie down the precise constants involved in the decay (i.e., $A_0 \neq A_1$). Third, the result (and its proof) does not directly exhibit information about the properties of an optimal ϵ-net. For a review of the theory see [66].

In this section we consider only the rough asymptotics of the ϵ-entropy via the critical exponent

$$\alpha^*(\mathcal{F}) = \sup\{\alpha\colon H_\epsilon(\mathcal{F}) = O(\epsilon^{-1/\alpha})\}.$$

If $H_\epsilon(\mathcal{F}) = \epsilon^{-1/\alpha}$ then $\alpha^*(\mathcal{F}) = \alpha$; but it is also true that if $H_\epsilon(\mathcal{F}) = \log(\epsilon^{-1})^\beta\cdot\epsilon^{-1/\alpha}$ then $\alpha^*(\mathcal{F}) = \alpha$. We should think of α^* as capturing only crude aspects of the asymptotics of ϵ-entropy, since it ignores "log terms" and related phenomena. We will show how to code in a way which achieves the rough asymptotic behavior.

B. Achieving the Exponent of ϵ-Entropy

Suppose we have a function ball $\mathcal{F}(\gamma)$ of a class F with unconditional basis (ϕ_k), and that, in addition, $\mathcal{F}(\gamma)$ has a certain tail compactness for the L^2 norm. In particular, suppose that in some fixed arrangement (k_i) of coordinates

$$\sup_{f\in\mathcal{F}(\gamma)} ||f - P_n(f;(\phi_k),(k_i))||_{L^2}^2 \leq C\cdot n^{-\mu} \tag{15.2}$$

for some $\mu > 0$.

We use nonlinear approximation in the basis (ϕ_k) to construct a binary coder giving an ϵ-description for all members f of $\mathcal{F}(\gamma)$. The idea: for an appropriate squared-L^2 distortion level ϵ^2, there is an $n = n(\epsilon, \mathcal{F})$ so that the best n-term approximation $Q_n(f;(\phi_k))$ achieves an approximation error $\leq \epsilon/2$. We then approximately represent f by digitally encoding the approximant. A reasonable way to do this is to concatenate a lossless binary code for the positions $k_1^*, \cdots, k_n^*$ of the important coefficients with a lossy binary code for quantized values $\tilde{a}_i$ of the coefficients a_i, taking care that the coefficients are encoded with an accuracy enabling accurate reconstruction of the approximant

$$||Q_n(f;(\phi_k)) - \sum_{i=1}^{n} \tilde{a}_i \phi_{k_i^*}||_2 \leq \epsilon/2. \tag{15.3}$$

Such a two-part code will produce a decodable binary representation having distortion $\leq \epsilon$ for every element of $\mathcal{F}(\gamma)$.

Here is a very crude procedure for choosing the quantization of the coefficients: there are n coefficients to be encoded; the coefficients can all be encoded with accuracy δ using a total of $\log(4/\delta)n$ bits. If we choose δ so that $n\delta^2 = \epsilon^2/4$, then we achieve (15.3).

To encode the positional information giving the locations $k_1^*, \cdots, k_n^*$ of the coefficients used in the approximant, use (15.2). This implies that there are C and ν so that, with $K_n = Cn^\nu$, for a given n, we can ignore coefficients outside the range $1 \leq k \leq K_n$. Hence, the positions can be encoded using at most $n \cdot \log_2(Cn^\nu)$ bits.

This gives a total codelength of $(C_1 + C_2 \log(n)) \cdot n$ bits. Now if $d_n^*(\mathcal{F}) \asymp n^{-m}$ then $n(\epsilon) \asymp \epsilon^{-1/m}$. Hence, assuming only that $d_n^*(\mathcal{F}) \asymp n^{-m}$ and that $\mathcal{F}$ is minimally tail compact, we conclude that

$$\alpha^*(\mathcal{F}) \leq m \tag{15.4}$$

and that a nonlinear approximation-based coder achieves this upper bound.

C. Lower Bound via $R(D)$

The upper bound realized through nonlinear approximation is sharp, as shown by the following argument [32]. Assume that $\mathcal{F}$ is a ball in a function class F that admits (ϕ_k) as an unconditional basis. The unconditional basis property means, roughly speaking, that $\mathcal{F}$ contains many very-high-dimensional hypercubes of appreciable sidelength. As we will see, $R(D)$ theory tells us precisely how many bits are required to D-faithfully represent objects in such hypercubes *chosen on average*. Obviously, one cannot faithfully represent *every object* in such hypercubes with fewer than the indicated number of bits.

Suppose now that (15.2) holds. We shall assume also that m is the best possible exponent; more precisely, we assume that, for some $A, B > 0$

$$A \cdot n^{-m} \leq d_n^*(\mathcal{F}, (\phi_k)) \leq B \cdot n^{-m}. \tag{15.5}$$

For fixed n, take f_n such that it attains the nonlinear n-width

$$||f_n - Q_n(f_n;(\phi_k))||_{L^2} = d_n^*(\mathcal{F}, (\phi_k)).$$

Let $\eta = \min_{1 \leq i \leq n} |a_i|$. By the unconditionality of the norm $||\theta||$, for every sequence of signs $(\pm_i: 1 \leq i \leq n)$ the object $g = \Sigma_i \pm_i \eta \phi_{k_i^*}$ belongs to $\mathcal{F}$. Hence we have constructed an orthogonal hypercube $\mathcal{H}$ of dimension n and sidelength η with $\mathcal{H} \subset \mathcal{F}$. Consider now h a random vertex of the hypercube $\mathcal{H}$. Representing this vertex with L^2 error $\leq n^{1/2} \cdot \eta \cdot \gamma$ is, owing to the orthogonality of the hypercube, the same as representing the sequence of signs with ℓ^2 error $\leq n^{1/2} \cdot \gamma$. Fix $\gamma < 1$; now use rate-distortion theory to obtain the rate-distortion curve for a binary discrete memoryless source with mean-squared error for single-letter difference distortion measure. No γ-faithful code for this source can use essentially fewer than $R(\gamma) \cdot n$ bits, where $R(\gamma) > 0$. Hence setting $\epsilon = n^{1/2} \cdot \eta \cdot \gamma$, we have

$$H_\epsilon(\mathcal{F}) \geq R(\gamma) \cdot n. \tag{15.6}$$

Since, for sufficiently large C, $d_{Cn}^*(\mathcal{F}, (\phi_k)) \leq A/2 \cdot n^{-m}$, we find that

$$||Q_n(f_n;(\phi_k)) - Q_{Cn}(f_n;(\phi_k))||_{L^2} \geq A/2 \cdot n^{-m}$$

which implies that $\eta^2(C-1)n \geq [A/2 \cdot n^{-m}]^2$, or $\eta \geq C' n^{-1/p}$. In other words,

$$\epsilon \geq \text{const}\, n^{1/2 - 1/p} = \text{const}\, n^{-m}.$$

We then conclude from (15.6) that

$$\alpha^*(\mathcal{F}) \geq m.$$

D. Order Asymptotics of ϵ-Entropy

The notion of achieving the exponent of the ϵ-entropy, while leading to very general results, is less precise than we would like. We now know that for a wide variety of classes of smooth functions, not only can the exponent be achieved, but also the correct order asymptotics (15.1) can be obtained by coders based on scalar quantization of the coefficients in a wavelet basis, followed by appropriate positional coding [14], [7], [15].

An important ingredient of nonlinear approximation results is the ability of the selected n-terms to occur at variable positions in the expansion. However, provision for the selected terms to occur in *completely arbitrary* arrangements of time-scale positions is actually not necessary, and we can obtain bit savings by exploiting the more limited range of possible positional combinations. A second factor is that most of the coefficients occurring in the quantized approximant involve small integer multiples of the quantum, and it can be known in advance that this is typically so at finer scales; provision for the "big coefficients" to occur in completely arbitrary orders among the significant ones is also not necessary. Consequently, there are further bit savings available there as well. By exploiting these two facts, one can develop coders which are within constant factors of the ϵ-entropy. We will explain this further in Section XVII below.

XVI. Comparison to Traditional Transform Coding

The results of the last few section offer an interesting comparison with traditional ideas of transform coding theory, for example the $R(D)$ theory of Section II.

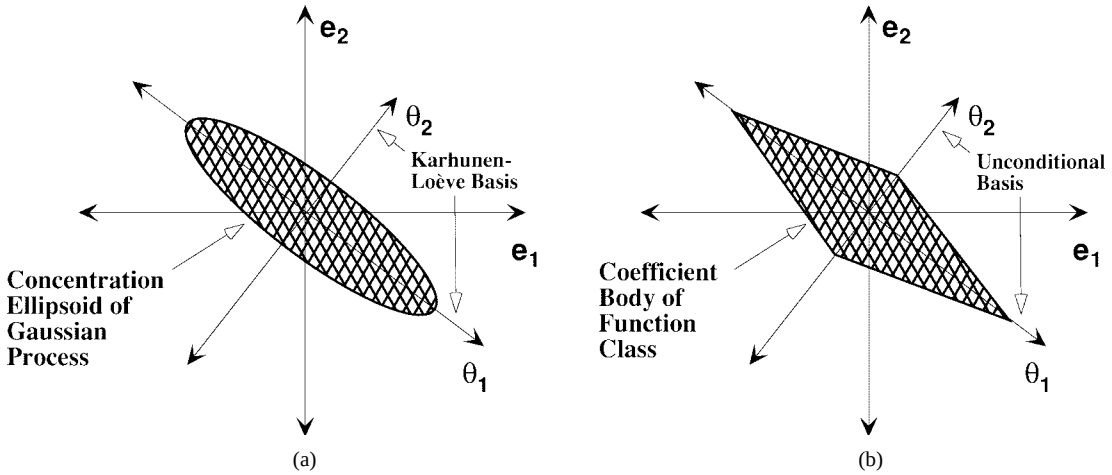

Fig. 6. A comparison of "diagonalization" problems. (a) Karhunen–Loève finds axes of symmetry of concentration ellipsoid. (b) Unconditional basis finds axes of symmetry of function class.

A. Nonlinear Approximation as Transform Coding

Effectively, the above results on nonlinear approximation can be interpreted as describing a very simple transform coding scheme. Accordingly, the results show that *transform coding in an unconditional basis for a class* F provides a near-optimal coding scheme for corresponding function balls $\mathcal{F}$.

Consider the following very simple transform coding idea. It has the following steps.

- We assume that we are given a certain *quantization step* q and a certain *bandlimit* K.
- Given an f to be compressed, we obtain the first K coefficients of f according to the system $(\phi_k): (a_k: 1 \leq k \leq K)$.
- We then apply simple scalar quantization to those coefficients, with stepsize q.

$$\tilde{a}_k = \operatorname{sgn}(a_k) \cdot q \cdot \operatorname{Floor}(|a_k|/q).$$

- Anticipating that the vast majority of these K quantized coefficients will turn out to be zero, we apply run-length coding to the quantized coefficients to efficiently code long runs of zeros.

This is a very simple coding procedure, perhaps even naive. But it turns out that

- if we apply this procedure to functions in the ball $\mathcal{F}(\gamma)$;
- if this ball arises from a function class F for which (ϕ_k) is an unconditional basis;
- if this ball obeys the minimal tail compactness condition (15.2), and the Stechkin numbers obey (15.5);
- if the coding parameters are appropriately calibrated;

the coder can achieve ϵ-descriptions of every $f \in \mathcal{F}(\gamma)$ achieving the rough asymptotics for the ϵ-entropy as described by $\alpha^*(\mathcal{F})$.

Essentially, the idea is that the behavior of the "nonlinear approximation" coder of Section XV can be realized by the "scalar quantization" coder, when calibrated with the right parameters. Indeed, it is obvious that the parameter K of the scalar quantizer and the parameter K_n of the nonlinear approximation coder play the same role; to calibrate the two coders they should simply be made equal. It then seems natural to calibrate q with n via

$$n(q) = \sum_k 1_{\{|\theta_k(f)| > q\}}.$$

After this calibration the two coders are roughly equivalent: they select approximants with precisely the same nonzero coefficients. The nonzero coefficients might be quantized slightly differently $(q \neq \delta)$, but using a few simple estimates deriving from the assumption $d_n^*(\mathcal{F}(\gamma)) \asymp n^{-m}$, one can see that they give comparable performance both in bits used and distortion achieved.

B. Comparison of Diagonalization Problems

Now we are in a position to compare the theory of Sections XI–XV with the Shannon $R(D)$ story of Section II. The $R(D)$ story of Section II says: to optimally compress a Gaussian stochastic process, transform the data into the Karhunen–Loève domain, and then quantize. The ϵ-entropy story of Sections XI–XV says: to (roughly) optimally compress a function ball $\mathcal{F}$, transform into the unconditional basis, and quantize.

In short, an orthogonal unconditional basis of a normed space plays the same role for function classes as the eigenbasis of the covariance plays for stochastic processes.

We have pointed out that an unconditional basis furnishes a generalization to nonquadratic forms of the concept of diagonalization of quadratic form. The Hölder class has, after an equivalent renorming, the shape of a hyperrectangle. The Bump Algebra has, after an equivalent renorming, the shape of an octahedron. A wavelet basis serves as axes of symmetry of these balls, just as an eigenbasis of a quadratic form serves as axes of symmetry of the corresponding ellipsoid.

For Gaussian random vectors, there is the concept of "ellipsoid of concentration;" this is an ellipsoidal solid which contains the bulk of the realizations of the Gaussian distribution. In effect, the Gaussian-$R(D)$ coding procedure identifies the problem of coding with one of transforming into the basis serving as axes of symmetry of the concentration ellipsoid. In comparison, our interpretation of the Kolmogorov-H_ϵ theory is that one should transform into the basis serving as axes of symmetry of the function ball, as illustrated by Fig. 6.

	Class	Process	Basis
L^2	Ellipsoid	Gaussian	Eigenbasis
non-L^2	Body	non-Gaussian	Unconditional Orthobasis

In effect, the function balls that we can describe by wavelet bases through norm equivalence, are often nonellipsoidal and in such cases do not correspond to the concentration ellipsoids of Gaussian phenomena. There is in some sense an analogy here to finding an *appropriate orthogonal transform for non-Gaussian data*. In the table at the top of this page, we record some aspects of this analogy without taking space to fully explain it. See also [34].

The article [27] explored using functional balls as models for real image data. The "Gaussian model" of data is an ellipsoid; but empirical work shows that, within the Besov and related scales, the nonellipsoidal cases provide a better fit to real image data. So the "better" diagonalization theory may well be the nontraditional one.

XVII. Beyond Orthogonal Bases: Trees

The effort of the harmonic analysis community to develop unconditional orthogonal bases can now be seen to have relevance to data compression and transform coding.

Unconditional bases exist only in very special settings, and the harmonic analysis community has developed many other interesting structures over the years—structures which go far beyond the concept of orthogonal basis.

We expect these broader notions of representation also to have significant data compression implications. In this section, we discuss representations based on the notion of dyadic tree and some data compression interpretations.

A. Recursive Dyadic Partitioning

A *recursive dyadic partition* (RDP) of a dyadic interval I_0 is any partition reachable by applying two rules.

- *Starting Rule:* $\{I_0\}$ itself is an RDP.
- *Dyadic Refinement:* If $\{I_1, \cdots, I_m\}$ is an RDP, and $I_j = I_{j,1} \cup I_{j,2}$ is a partition of the dyadic interval I_j into its left and right dyadic subintervals, then $\{I_1, \cdots, I_{j-1}, I_{j,1}, I_{j,2}, I_{j+1}, \cdots, I_m\}$ is a new RDP.

RDP's are also naturally associated to binary trees; if we label the root node of a binary tree by I_0 and let the two children of the node correspond to the two dyadic subintervals of I_0, we associate with each RDP a tree whose terminal nodes are subintervals comprising members of the partition. This correspondence allows us to speak of methods exploiting RDP's as "tree-structured methods."

Recursive dyadic partitioning has played an important role throughout the subject of harmonic analysis, as one can see from many examples in [89] and [45]. It is associated with ideas like the Whitney decomposition of the 1930's, and the Calderón–Zygmund Lemma in the 1950's.

A powerful strategy in harmonic analysis is to construct RDP's according to "stopping time rules" which describe when to stop refining. This gives rise to data structures that are highly adapted to the underlying objects driving the construction. One then obtains analytic information about the objects of interest by combining information about the structure of the constructed partition and the rule which generated it. In short, recursive dyadic partitioning is a flexible general strategy for certain kinds of delicate nonlinear analysis.

The RDP concept allows a useful decoration of the time-scale plane, based on the family of time-scale rectangles $R(I)$ which we introduced in Section XII-B. If we think of all the intervals visited in the sequential construction of an RDP starting from the root I_0 as intervals where something "important is happening" and the ones not visited, i.e., those occurring at finer scales than intervals in the partition, as ones where "not much is happening," we thereby obtain an adaptive labeling of the time-scale plane by "importance."

Here is a simple example, useful below. Suppose we construct an RDP using the rule "stop when no subinterval of the current interval has a wavelet coefficient $|\theta_I| > \epsilon$;" the corresponding labeling of the time-scale plane shows a "stopping time region" outside of which all intervals are unimportant, i.e., are associated to wavelet coefficients $\leq \epsilon$. For later use, we call this stopping-time region the *hereditary cover of the set of wavelet coefficients larger than* ϵ; it includes not only the cells $R(I)$ associated to "big" wavelet coefficients, but also the ancestors of those cells. The typical appearance of such a region, in analyzing an object with discontinuities, is that of a region with very fine "tendrils" reaching down to fine scales in the vicinity of the discontinuities; the visual appearance can be quite striking; see Fig. 7.

Often, as in the Whitney and Calderón–Zygmund constructions, one runs the "stopping time" construction only once, but there are occasions where running it repeatedly is important. Doing so will produce a sequence of nested sets; in the hereditary cover example, one can see that running the stopping time argument for $\epsilon = 2^{-k}$ will give a sequence of regions; outside of the kth one, no coefficient can be larger than 2^{-k}.

In the 1950's and early 1960's, Carleson studied problems of interpolation in the upper half-plane. In this problem, we suppose we are given prescribed values u_k at an irregular set of points (t_k, y_k) in the upper half-plane

$$U(t_k, y_k) = u_k, \qquad k = 1, 2, \cdots$$

and we ask whether there is a bounded function f on the line whose Poisson integral U obeys the stated conditions. Owing to the connection with wavelet transforms, this is much like asking whether, from a given scattered collection of data

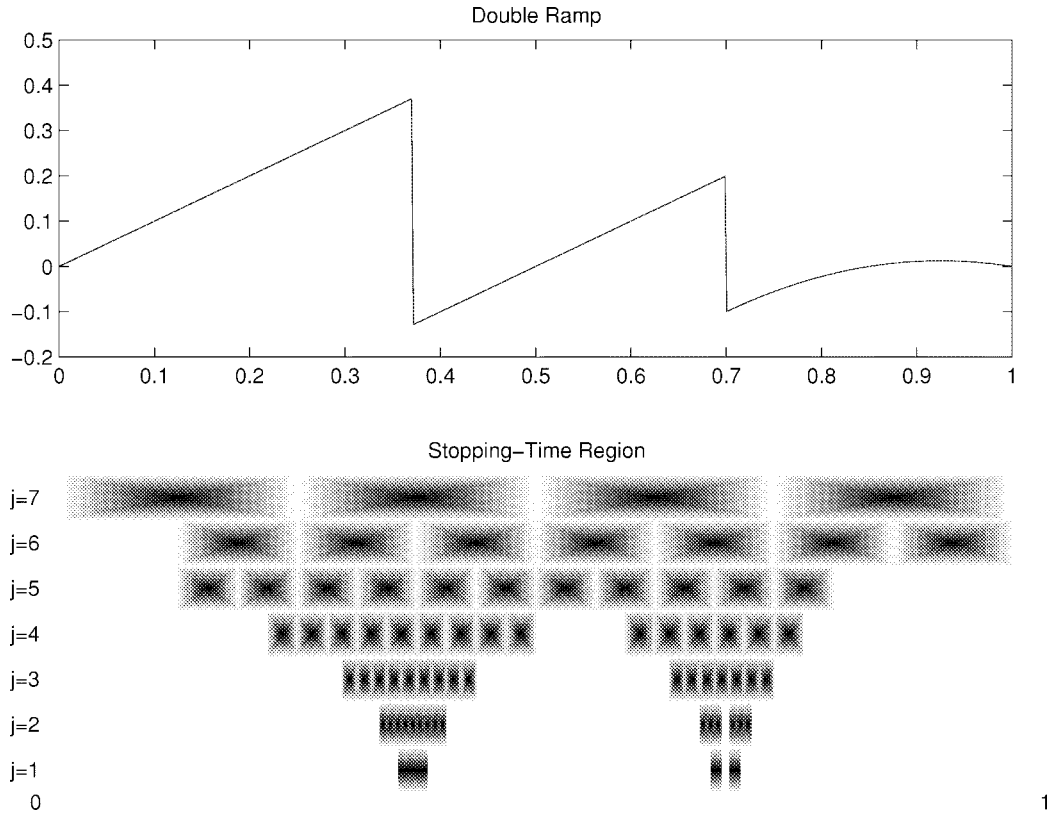

Fig. 7. An object, and the stopping-time region for the q-big coefficients.

about the wavelet transform (b_k, a_k) we can reconstruct the underlying function. Carleson developed complete answers to this problem, through a method now called the "Corona Construction," based on running stopping time estimates repeatedly. Recently P. Jones [57] discussed a few far-reaching applications of this idea in various branches of analysis, reaching the conclusion that it "... is one of the most malleable tools available."

B. Trees and Data Compression

In the 1960's, Birman and Solomjak [8] showed how to use recursive dyadic partitions to develop coders achieving the correct order of Kolmogorov ϵ-entropy for L^p-Sobolev function classes.

The Birman–Solomjak procedure starts from a function f of interest, and constructs a partition based on a parameter δ. Beginning with I_0 the whole domain of interest, the stopping rule refines the dyadic interval I only if approximating the function f on I by a polynomial of degree m gives an error exceeding δ. Call the resulting RDP a δ-partition. The coding method is to construct a δ-partition, where the parameter δ is chosen to give a certain desired global accuracy of approximation, and to represent the underlying function by digitizing the approximating polynomials associated with the δ-partition. By analyzing the number of pieces of an ϵ-partition, and the total approximation error of a δ-partition, and by studying the required number of bits to code for the approximating polynomials on each individual, they showed that their method gave optimal order bounds on the number of bits required to ϵ-approximate a function in a $W_p^m(\gamma)$ ball.

This type of result works in a broader scale of settings than just the L^p-Sobolev balls. A modern understanding of the underlying nonlinear approximation properties of this approach is developed in [28].

This approximation scheme is highly adaptive, with the partition adapting spatially to the structure of the underlying object. The ability of the scheme to refine more aggressively in certain parts of the domain than in others is actually necessary to obtaining optimal order bounds for certain choices of the index p defining the Sobolev norm; if we use L^2 norm for measuring approximation error, the adaptive refinement is necessary whenever $p<2$. Such spatially inhomogeneous refinement allows to use fewer bits to code in areas of unremarkable behavior and more bits in areas of rapid change.

RDP ideas can also be used to develop wavelet-domain transform coders that achieve the optimal order ϵ-entropy bounds over L^p-Sobolev balls $W_p^m(\gamma)$, with $m+1/2-1/p>0$. Recall that the argument of Section XV showed that any coding of such a space must require $\geq C\epsilon^{-1/m}$ bits, while a very crude transform coding was offered that required order $\log(\epsilon^{-1})\epsilon^{-1/m}$ bits. We are claiming that RDP ideas can be used to eliminate the log term in this estimate [14]. In effect, we are saying that by exploiting trees, we can make the cost of coding the position of the big wavelet coefficients at worst comparable to the ϵ-entropy.

The key point about a function in L^p-Sobolev space is that, on average, the coefficients $\theta_{j,k}$ decay with decreasing scale 2^{-j}. Indeed, from the g-function representation described in Section XII-D, we can see that the wavelet coefficients of

such a function obey, at level j

$$\left(\sum_k |\theta_{j,k}|^p\right)^{1/p} \le \text{Const} \cdot 2^{-j(m+1/2-1/p)}. \qquad (17.1)$$

In the range $m+1/2-1/p>0$, this inequality controls rather powerfully the number and position of nonzero coefficients in a scalar quantization of the wavelet coefficients. Using it in conjunction with RDP ideas allows to code the position of the nonzero wavelet coefficients with far fewer bits than we have employed in the run-length coding ideas of Section XV.

After running a scalar quantizer on the wavelet coefficients, we are left with a scattered collection of nonzero quantizer outputs; these correspond to coefficients exceeding the quantizer interval q. Now form the hereditary cover of the set of all coefficients exceeding q, using the stopping time argument described in Section XVII-A. The coder will consist of two pieces: a lossless code describing this stopping time region (also the positions within the region containing nonzero quantizer outputs), and a linear array giving the quantizer outputs associated with nonzero cells in the region.

This coder allows the exact same reconstruction that was employed using the much cruder coder of Section XV. We now turn to estimates of the number of bits required.

A stopping-time region can be easily coded by recording the bits of the individual refine/don't refine decisions; we should record also the presence/absence of a "big" coefficient at each nonterminal cell in the region, which involves a second array of bits.

The number of bits required to code for a stopping-time region is, of course, proportional to the number of cells in the region itself. We will use in a moment the inequality (17.1) to argue that in a certain maximal sense the number of cells in a hereditary cover is not essentially larger than the original number of entries generating the cover. It will follow from this that the number of bits required to code the positions of the nonzero quantizer outputs is in a maximal sense proportional to the number n of nonzero quantizer outputs. This, in turn, is the desired estimate; the cruder approach of Section XV gave only the estimate $n \log(n)$, where n is the number of quantizer outputs. For a careful spelling-out of the kind of argument we are about to give, see [33, Lemma 10.3].

We now estimate the maximal number of cells in the stopping-time region, making some remarks about the kind of control exerted by (17.1) on the arrangement and sizes of the wavelet coefficients; compare Fig. 7. The wavelet coefficients of functions in an L^p-Sobolev ball associated to I_0 have a "Last Full Level" J_0: this is the finest level j for which one can find some f in the ball such that for all $I \subset I_0$ of length 2^{-j}, the associated wavelet coefficient exceeds q in absolute value. By (17.1), any such j obeys $j \le j_0$, where j_0 is the real solution of

$$2^{j_0/p} q = C \cdot 2^{-j_0(\alpha+1/2-1/p)}.$$

The wavelet coefficients also have a "First Empty Level" J_1, i.e., a coarsest level beyond which all wavelet coefficients are bounded above in absolute value by q. By (17.1), any nonzero quantizer output occurs at $j \le j_1$, where

$$q = C \cdot 2^{-j_1(\alpha+1/2-1/p)}.$$

At scales intermediate between the two special values, $j_0 < j < j_1$, there are possibly "sparse levels" which can contain wavelet coefficients exceeding q, in a subset of the positions. However, the maximum possible number n_j of such nonzero quantizer outputs at level j thins out rapidly, in fact exponentially, with increasing $j - j_0$. Indeed, we must have

$$n_j^{1/p} q \le C \cdot 2^{-j(\alpha+1/2-1/p)}$$

and from $n_{\lfloor j_0 \rfloor} \le 2^{j_0}$ we get

$$n_j \le \text{Const} \cdot 2^{j_0} \cdot 2^{-\beta(j-j_0)}$$

with $\beta > 0$.

In short, there are about 2^{j_0} possible "big" coefficients at the level nearest j_0, but the maximal number of nonzero coefficients decays exponentially fast at scales away from j_0. In an obvious probabilistic interpretation of these facts the "expected distance" of a nonzero coefficient below j_0 is $O(1)$.

Now obviously, in forming the hereditary cover of the positions of nonzero quantizer outputs, we will not obtain a set larger than the set we would get by including all positions through level j_0, and also including all "tendrils" reaching down to positions of the nonzeros at finer scales than this. The expected number of cells in a tendril is $O(1)$ and the number of tendrils is $O(2^{j_0})$. Therefore, the maximal number of cells in the q-big region is not more than $C 2^{j_0(q)}$. The maximal number of nonzeros is of size $C 2^{j_0(q)}$.

These bounds imply the required estimates on the number of bits to code for positional information.

One needs to spend a little care also on the encoding of the coefficients themselves. A naive procedure would spend $O(\log(\epsilon^{-1}))$ on each of $O(\epsilon^{-1/m})$ coefficients, leading to a crude estimate requiring $\log(\epsilon^{-1})\epsilon^{-1/m}$ bits, now to encode the coefficient information. By making use of the nested sets in the hereditary cover described earlier, one can structure the coefficients to be retained into layers, in which the label of the layer indicates how many (or few) bits need to be spent on coefficients in that layer. One can estimate, similarly to what was done above, that the layers with large coefficients, for which more bits are required, contain few elements; layers corresponding to much smaller coefficients have many more elements, but because their label already restricted their size, we spend fewer bits on them to specify them with the same precision. Accounting for the cost in bits of this procedure, one finds that encoding the coefficients also requires $O(\epsilon^{-1/m})$ bits only.

This encoding strategy can be modified so as to be progressive and universal. By adding one bit for each existing coefficient and new bits to specify the new coefficients and their position, it becomes progressive. Each additional stream of bits serves to add new detail to the existing approximation in a nonredundant way. The encoding is also universal in the sense that the encoder does not need to know the characteristics of the class $\mathcal{F}$: the encoder is defined once and for all and enjoys the property that each class $\mathcal{F}$ which has Φ as

an unconditional basis will be optimally encoded with respect to Kolmogorov entropy. It is interesting to note that practical wavelet-based encoders, such as those introduced for images in [86] and [82], carry out a similar procedure.

XVIII. Beyond Time Scale: Time Frequency

The analysis methods described so far—wavelets or trees—were *Time-Scale* methods. These methods associate, to a simple function of time only, an object extending over both time and scale, and they extract analytic information from that two-variable object.

An important complement to this is the family of *Time-Frequency* methods. Broadly speaking, these methods try to identify the different frequencies present in an object at time t.

In a sense, time-scale methods are useful for compressing objects which display punctuated smoothness—e.g., which are typically smooth away from singularities. Time-frequency methods work for oscillatory objects where the frequency content changes gradually in time.

One thinks of time-scale methods as more naturally adapted to image data, while time-frequency methods seem more suited for acoustic phenomena. For some viewpoints on mathematical time-frequency analysis, see [21], [23], and [39].

In this section we very briefly cover time-frequency ideas, to illustrate some of the potential of harmonic analysis in this setting.

A. Modulation Spaces

Let $g(t) = \sqrt{2} \cdot \pi^{1/2} \cdot e^{-t^2/2}$ be a Gaussian window. Let $g_{u,\omega}(t) = \exp\{-i\omega t\} g(t-u)$ denote the *Gabor function* localized near time u and frequency ω. The family of all Gabor functions provides a range of oscillatory behavior associated with a range of different intervals in time. These could either be called "time-frequency atoms" or, more provocatively, "musical notes."

Let S denote the collection of all functions $f(t)$ which can be decomposed as

$$f = \sum_{i=1}^{\infty} a_i g_{u_i,\omega_i}(t)$$

where $\Sigma_i |a_i| < \infty$. Let $||f||_S$ denote the smallest value $\Sigma_i |a_i|$ occurring in any such decomposition. This provides a normed linear space of functions, which Feichtinger calls the *Segal algebra* [40]. For a given amplitude A, the norm constraint $||f||_S \le \gamma$ controls the number of "musical notes" of strength A which f can contain: $n \cdot A \le \gamma$. So such a norm controls in a way the complexity of the harmonic structure of f.

A special feature of the above class is that there is no unique way to obtain a decomposition of the desired type, and no procedure is specified for doing so. As with our earlier analyses, it would seem natural to seek a basis for this class; one could even ask for an unconditional basis.

This Segal algebra is an instance of Feichtinger's general family *modulation spaces* $M^{\alpha}_{p,q}(\boldsymbol{R})$ and we could more generally ask for unconditional bases for any of these spaces. The modulation spaces offer an interesting scale of spaces in certain respects analogous to L^p-Sobolev and related Besov and Triebel scales; in this brief survey, we are unable to describe them in detail.

B. Continuous Gabor Transform

Even older than the wavelet transform is the continuous Gabor transform (CGT). The Gabor transform can be written in a form analogous to (12.8) and (12.10)

$$(Gf)(u,\omega) = \langle f, g_{u,\omega} \rangle \tag{18.1}$$

$$f(t) = \int\int (Gf)(u,\omega) g_{u,\omega}\, du\, d\omega. \tag{18.2}$$

Here $g_{u,\omega}$ can be a Gabor function as introduced above; it can also be a "Gabor-like" function generated from a non-Gaussian window function g.

If both g and its Fourier transform $\hat{g}$ are "concentrated" around 0 (which is true for the Gaussian), then (18.1) captures information in f that is localized in time (around u) and frequency (around ω); (18.2) writes f as a superposition of all the different time-frequency pieces.

Paternity of this transform is not uniquely assignable; it may be viewed as a special case in the theory of square-integrable group representations, a branch of abstract harmonic analysis [47]; it may be viewed as a special case of decomposition into canonical coherent states, a branch of mathematical physics [59]. In the signal analysis literature it is generally called the CGT to honor Gabor [44], and that appellation is very fitting on this occasion. Gabor proposed a "theory of communication" in 1946, two years before Shannon's work, that introduced the concept of logons—information carrying "cells"—highly relevant to the present occasion, and to Shannon's sampling theory.

The information in G is highly redundant. For instance, $Gf(u,\omega)$ captures what happens in f not only at time $t = u$, but also t near u and similarly in frequency. The degree to which this "spreading" occurs is determined by the choice of the window function g: if g is very narrowly concentrated around $t = 0$, then the sensitivity of $Gf(u,\omega)$ to the behavior of $f(t)$ is concentrated to the vicinity of $t = u$. The Heisenberg Uncertainty Principle links the concentration of g with that of the Fourier transform $\hat{g}$. If we define

$$\Delta_t^2(g) = \int t^2 |g|^2(t)\, dt$$

and

$$\Delta_\omega^2(g) = \int \omega^2 |\hat{g}|^2(\omega)\, d\omega$$

then $\Delta_t \cdot \Delta_\omega \ge \sqrt{\frac{\pi}{2}}\, ||g||^2_{L^2}$, so that as g becomes more concentrated, $\hat{g}$ becomes less so, implying that the frequency sensitivity of $Gf(u,\omega)$ becomes more diffuse when we improve the time localization of g. Formalizing this, let $\rho_g(u,\omega)$ denote the rectangle of dimensions $\Delta_t(g) \times \Delta_\omega(g)$ centered at (u,ω). The choice of g influences the shape of the region (more narrow in time, but elongated in frequency if we choose g very concentrated around 0); the area of the cell is bounded below by the Uncertainty Principle. We think of each point $Gf(u,\omega)$ as measuring properties of a "cell" in the time-frequency domain, indicating the region that is "captured" in

$Gf(u,\omega)$. For example, if $\rho_g(u,\omega)$ and $\rho_g(u',\omega')$ are disjoint, we think of the corresponding Gf values as measuring disjoint properties of f.

C. Atomic Decomposition

Together, (18.1) and (18.2) can be written as

$$\iint \langle \cdot, g_{u,\omega}\rangle g_{u,\omega}\, du\, d\omega = Id, \tag{18.3}$$

another resolution of the identity operator.

For parameters δ_t, δ_ω to be chosen, define equispaced time points $u_k = k\delta_t$ and frequency points $\omega_\ell = \ell\delta_\omega$. Consider now a family of time-frequency rectangles

$$R_{k,\ell} = \{(u,\omega)\colon |u - u_k| \le \delta_t/2; |\omega - \omega_\ell| \le \delta_\omega/2\}.$$

Evidently, these rectangles are disjoint and tile the time-frequency plane.

Proceeding purely formally, we can partition the integral in the resolution of the identity to get a decomposition $Id = \Sigma_{k,\ell} A_{k,\ell}$ with individual operators

$$A_{k,\ell} = \iint_{R_{k,\ell}} \langle \cdot, g_{u,\omega}\rangle g_{u,\omega}\, du\, d\omega.$$

This provides formally an atomic decomposition

$$f = \sum_{k,\ell} A_{k,\ell} f. \tag{18.4}$$

In order to justify this approach, we would have to justify treating a rectangle $R_{k,\ell}$ as a single coherent region of the time-frequency plane. Heuristically, this coherence will apply if δ_t is smaller than the spread $\Delta_t(g)$ and if δ_ω is smaller than the spread $\Delta_\omega(g)$ of $\hat{g}$. In short, if the rectangle $R_{k,l}$ has the geometry of a Heisenberg cell, or smaller, then the above approach makes logical sense.

One application for atomic decomposition would be to characterize membership in the Segal Algebra S. Feichtinger has proved that, with g a sufficiently nice window, like the Gaussian, if we pick δ_t and δ_ω sufficiently small, atomic decomposition allows to measure the norm of S. Let $s_{k,l} = \int_{R_{k,\ell}} |G(u,\omega)|$ measure the size of the atom $A_{k,l}f$; then a function f is in S if and only if $\Sigma_{k,l}\, s_{k,l} < \infty$. Moreover, $||A_{k,\ell}f||_S \le \text{Const} \cdot s_{k,l}$ so the series $f = \Sigma_{k,\ell} A_{k,\ell} f$ represents a decomposition of f into elements of S, and

$$||f||_S \asymp \sum_{k,l} s_{k,l}.$$

So we have an equivalent norm for the S-norm. This in some sense justifies the Gabor "logon" picture, as it shows that an object can really be represented in terms of elementary pieces, those pieces being associated with rectangular time-frequency cells, and each piece uniformly in S.

D. Orthobasis

Naturally, one expects to obtain not just an atomic decomposition, but actually an orthobasis. In fact, Gabor believed that one could do so. One approach is to search for a window g, not necessarily the Gaussian, and a precise choice of δ_u and δ_ω so that the samples $Gf(u_k,\omega_l)$ at equispaced points $u_k = k\delta_t\omega_\ell = \ell\delta_\omega$ provide an exact norm equivalence

$$\sum_{k,\ell} |Gf(u_k,\omega_l)|^2 = \int |f(t)|^2\, dt.$$

While this is indeed possible, a famous result due to Balian and Low shows that it is not possible to achieve orthogonality using a g which is nicely concentrated in both time and frequency; to get an orthogonal basis from Gabor functions requires to have $\Delta_t(g) \cdot \Delta_\omega(g) = +\infty$. Hence the geometric picture of localized contributions from rectangular regions is not compatible with an orthogonal decomposition. A related effect is that the resulting Gabor orthogonal basis would not provide an unconditional bases for a wide range of modulation spaces. In fact, for certain modulation spaces, nonlinear approximation in such a basis would not behave optimally; e.g., we have examples of function balls $\mathcal{F}$ in modulation spaces where $d_n^*(\mathcal{F}) \ll d_n(\mathcal{F}, \text{GABOR ORTHOBASIS})$.

A way out was discovered with the construction of so-called *Wilson* orthobases for L^2 [22]. These are Gabor-like bases, built using a special smooth window $g(t)$ of rapid decay, and consist of basis elements $\phi_{k,\ell}$, where k is a position index and ℓ is a frequency index. k runs through the integers, and ℓ runs through the nonnegative integers; the two parameters vary independently, except that $\ell = 0$ is allowed in conjunction only with even k. In detail

$$\phi_{k,l}(t) = \begin{cases} \sqrt{2} g(t - k2\pi)\cos\left(\frac{\ell}{2}t\right), & k = 0, \pm 2, \pm 4, \cdots \\ & \ell = 1, 2, 3, \cdots \\ g(t - k2\pi), & k = 0, \pm 2, \pm 4, \cdots \\ & \ell = 0 \\ \sqrt{2} g(t - k2\pi)\sin\left(\frac{\ell}{2}t\right), & k = \pm 1, \pm 3, \cdots \\ & \ell = 1, 2, 3, \cdots . \end{cases}$$

Owing to the presence of the cosine and sine terms, as opposed to complex exponentials, the $\phi_{k,\ell}$ are not truly Gabor functions; but they can be viewed as superpositions of certain pairs of Gabor functions. The Gabor functions used in those pairs do not fill out the vertices of a single rectangular lattice, but instead they use a subset of the vertices of two distinct lattices. Hence the information in the Wilson coefficients derives indeed from sampling of the CGT with special generating window $g(t)$, only the samples must be taken on two interleaved Cartesian grids; and the samples must be combined in pairs in order to create the Wilson coefficients, as shown in Fig. 8.

The resulting orthobasis has good analytic properties. Grochenig and Walnut [51] have proved that it offers an unconditional basis for all the modulation spaces; in particular, it is an unconditional basis for S. As a result, Wilson bases are best orthobases for nonlinear approximation; for function balls $\mathcal{F}$ arising from a wide range of modulation spaces

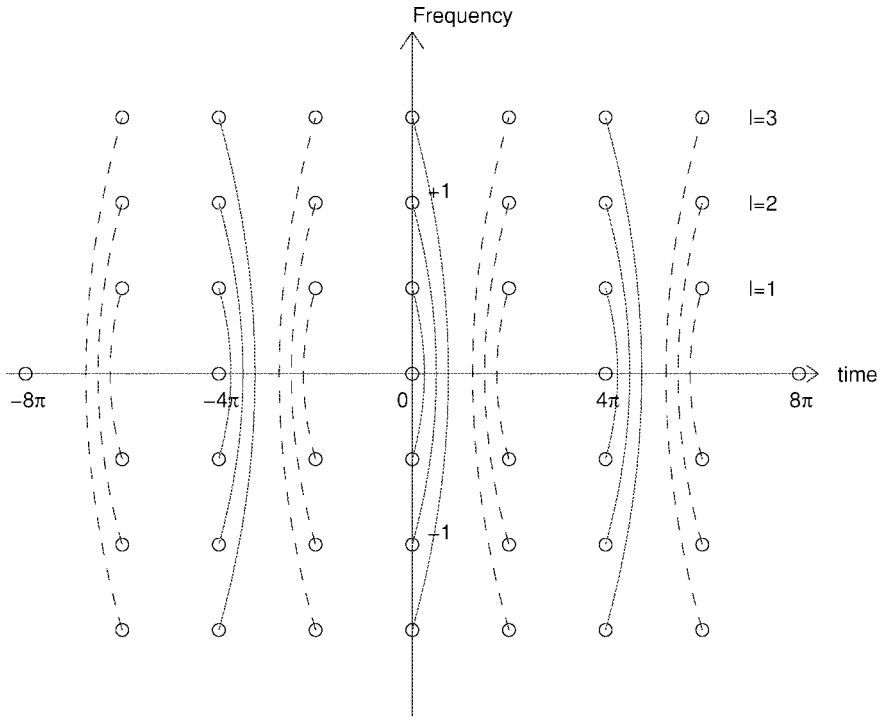

Fig. 8. The sampling set associated with the Wilson basis. Sampling points linked by solid lines are combined by addition. Sampling points linked by dashed lines are combined by subtraction.

$d_n^*(\mathcal{F}) \asymp d_n(\mathcal{F}, \text{WILSON ORTHOBASIS})$. See also [50]. There are related data compression implications, although to state them requires some technicalities, for example, imposing on the objects to be compressed some additional decay conditions in both the time and frequency domains.

E. Adaptive Time-Frequency Decompositions

In a sense, CGT analysis/Gabor Atomic Decomposition/ Wilson bases merely scratch the surface of what one hopes to achieve in the time-frequency domain. The essential limitation of these tools is that they are *monoresolution*. They derive from a uniform rectangular sampling of time and frequency which is suitable for some phenomena; but it is easy to come up with models requiring highly nonuniform approaches.

Define the multiscale Gabor dictionary, consisting of Gabor functions with an additional scale parameter δ

$$g_{(u,\omega,\delta)}(t) = \exp\{i\omega(t-u)\}\exp\{(t-u)^2/\delta^2\}.$$

For $0 < p \leq 1$, consider the class MS^p of objects f having a multiscale decomposition

$$f = \sum_i a_i g_{(u_i,\omega_i,\delta_i)} \tag{18.5}$$

with $\Sigma |a_i|^p < C^p$. In a sense, this class is obtained by combining features of the Bump Algebra—multiscale decomposition—and the Segal Algebra—time-frequency decomposition. This innocent-looking combination of features responds to a simple urge for common-sense generalization, but it gets us immediately into mathematical hot water. Indeed, one cannot use a simple monoscale atomic decomposition based on the CGT to effectively decompose such an f into its pieces at different scales; one cannot from the monoscale analysis infer the size of the minimal C appearing in such a decomposition. Moreover, we are unaware of any effective means of decomposing members of this class in a way which achieves a near-optimal representation, i.e., a representation (18.5) with near-minimal C.

One might imagine performing a kind of "multiscale Gabor transform"—with parameters time, scale, and frequency, and developing an atomic decomposition or even a basis based on a three-dimensional rectangular partitioning of this time-scale-frequency domain, but this cannot work without extra precautions [95]. The kind of sampling theorem and norm equivalence result one might hope for has been proven impossible in that setting.

An important organizational tool in getting an understanding of the situation is to consider dyadic Heisenberg cells only, and to understand the decompositions of time and frequency which can be based upon them. These dyadic Heisenberg cells $H(j,k_1,k_2)$ are of side $2^{-j} \times 2^j$ and volume 1, with lower left corner at $k_1/2^j, k_2 2^j$ in the time-frequency plane. The difficulty of the time-frequency-scale decomposition problem is expressed by the fact that each Heisenberg cell overlaps with infinitely many others, corresponding to different aspect ratios at the same location. This means that even in a natural discretization of the underlying parameter space, there are a wide range of multiscale Gabor functions interacting with each other at each point (u,ω) of the time-frequency domain.

This kind of dyadic structuring has been basic to the architecture of some key results in classical analysis, namely, Carleson's proof of the a.e. convergence of Fourier series, and also Fefferman's alternate proof of the Carleson theorem. Both of those proofs were based on dyadic decomposition

of time and frequency, and the driving idea in those proofs is to find ingenious ways to effectively combine estimates from all different Heisenberg cells despite the rather massive degree of overlap present in the family of all these cells. For example, Fefferman introduced a tree-ordering of dyadic Heisenberg cells, and was able to get effective estimates by constructing many partitions of the time-frequency domain according to the properties of the Fourier partial sums he was studying, obtaining a decomposition of the sums into operators associated with special time-frequency regions called trees, and combining estimates across trees into forests.

Inspired by their success, one might imagine that there is some way to untangle all the overlap in the dyadic Heisenberg cells and develop an effective multiscale time-frequency analysis. The goal would be to somehow organize the information coming from dyadic Heisenberg cells to iteratively extract from this time-scale-frequency space various "layers" of information. As we have put it this program is, of course, rather vague.

More concrete is the idea of *nonuniform tiling of the time-frequency plane*. Instead of decomposing the plane by a sequence of disjoint congruent rectangles, one uses rectangles of various shapes and sizes, where the cells have been chosen specially to adapt to the underlying features of the object being considered.

The idea of adaptive tiling is quite natural in connection with the problem of multiscale time-frequency decomposition, though it originally arose in other areas of analysis. Fefferman's survey [41] gives examples of this idea in action in the study of partial differential equations, where an adaptive time-frequency tiling is used to decompose operator kernels for the purpose of estimating eigenvalues.

Here is how one might expect adaptive tiling to work in a multiscale time-frequency setting. Suppose that an object had a representation (18.5) where the fine-scale atoms occurred at particular time locations well separated from the coarse-scale atoms. Then one could imagine constructing a time-frequency tiling that had homologous structure: rectangles with finer time scales where the underlying atoms in the optimal decomposition needed to be fine-scale; rectangles with coarser time-scale elsewhere. The idea requires more thought than it might at first seem, since the rectangles cannot be chosen freely; they must obey the Heisenberg constraint. The dyadic Heisenberg system provides a constrained set of building blocks which often can be fit together quite easily to create rather inhomogeneous tilings. Also, the rectangles must correspond rigorously to an analyzing tool: there must be an underlying time-frequency analysis with a width that is changing with spatial location. It would be especially nice if one could do this in a way providing true orthogonality.

In any event, it is clear that an appropriate multiscale time-frequency analysis in the setting of MS^p or related classes cannot be constructed within a single orthobasis. At the very least, one would expect to consider large families of orthobases, and select from such a family an individual basis best adapted for each individual object of interest. However, even there it is quite unlikely that one could obtain true characterization of a space like MS^p; i.e., it is unlikely that even within the broader situation of adaptively chosen orthobases, the decompositions one could obtain would rival an optimal decomposition of the form (18.5), i.e., a decomposition with minimal C. A corollary to this is that we really know of no effective method for data compression for this class: *there is no effective transform coding for multiscale time-frequency classes*.

XIX. Computational Harmonic Analysis

The theoretical constructions of harmonic analysts correspond with practical signal-processing tools to a remarkable degree. The ideas which are so useful in the functional viewpoint, where one is analyzing functions $f(t)$ of a continuous argument, correspond to completely analogous tools for the analysis of discrete-time signals $x(n)$. Moreover, the tools can be realized as fast algorithms, so that on signals of length N they take order N or $N \log(N)$ operations to complete. Thus corresponding to the theoretical developments traced above, we have today fast wavelet transforms, fast Gabor transforms, fast tree approximation algorithms, and even fast algorithms for adapted multiscale time-frequency analysis.

The correspondence between theoretical harmonic analysis and effective signal processing algorithms has its roots in two specific facts which imply a rather exact connection between the functional viewpoint of this paper and the digital viewpoint common in signal processing.

- *Fast Fourier Transform:* The finite Fourier transform provides an orthogonal transform for discrete-time sequences which, in a certain sense, matches perfectly with the classical Fourier series for functions on the circle. For example, on appropriate trigonometric polynomials, the first N Fourier series coefficients are (after normalization) precisely the N finite Fourier coefficients of the digital signal obtained by sampling the trigonometric polynomial. This fact provides a powerful tool to connect concepts from the functional setting with the discrete-time signal setting. The availability of fast algorithms has made this correspondence a computationally effective matter. It is an eerie coincidence that the most popular form of the FFT algorithm prefers to operate on signals of dyadic length; for connecting theoretical harmonic analysis with the digital signal processing domain, the dyadic length is also the most natural.
- *Sampling Theorem:* The classical Shannon sampling theorem for bandlimited functions on the line has a perfect analog for digital signals obeying discrete-time bandlimiting relations. This is usually interpreted as saying that one can simply subsample a data series and extract the minimal nonredundant subset. A different way to put it is that there is an orthonormal basis for the bandlimited signals and that sampling provides a fast algorithm for obtaining the coefficients of a signal in that basis.

There are, in addition, two particular tools for partitioning signal domains which allow effective digital implementation of breaking a signal into time-scale or time-frequency pieces.

- *Smooth Orthonormal Localization:* Suppose one takes a discrete-time signal of length $2N$ and breaks it into two

subsignals of length N, corresponding to the first half and the second half. Now subject those two halves to further processing, expanding each half into an orthonormal basis for digital signals of length N. In effect, one is expanding the original length $2N$ signal into an orthonormal basis for length $2N$. The implicit basis functions may not be well-behaved near the midpoint. For example, if the length N basis is the finite Fourier basis, then the length $2N$ basis functions will be discontinuous at the segmentation point. This discontinuity can be avoided by changing the original "splitting into two pieces" into a more sophisticated partitioning operator based on a kind of smooth orthonormal windowing. This involves treating the data near the segmentation point specially, taking pairs of values located equal distances from the segmentation point and on opposite sides and, instead of simply putting one value in one segment and the other value in the other segment, one puts special pairs of linear combinations of the two values in the two halves; see for example [71], [72], and [3].

- *Subband Partitioning:* Suppose we take a discrete-time signal, transform it into the frequency domain, and break the Fourier transform into two pieces, the high and low frequencies. Now transform the pieces back into the time domain. As the pieces are now bandlimited/bandpass, they can be subsampled, creating two new vectors consisting of the 'high frequency' and 'low frequency' operations. The two new vectors are related to the original signal by orthogonal transformation, so the process is in some sense exact. Unfortunately, the brutal separation into high and low frequencies has many undesirable effects. One solution to this problem would have been to apply smooth orthonormal localization in the frequency domain. A different approach, with many advantages, is based on the time domain method of *conjugate quadrature filters*.

  In this approach, one applies a special pair of digital filters, high- and lowpass, to a digital signal of length $2N$, and then subsamples each of the two results by a factor two [88], [76], [97], [99]. The result is two signals of length N, so that the original cardinality of $2N$ is preserved, and *if the filters are very specially chosen*, the transform can be made orthogonal. The key point is that the filters can be short. The most elementary example is to use a highpass filter with coefficients $(1/\sqrt{2}, -1/\sqrt{2})$ and a lowpass filter with coefficients $(1/\sqrt{2}, 1/\sqrt{2})$. The shortness of this filter means that the operator does not have the time-localization problems of the frequency-domain algorithm, but unfortunately this filter pair will not have very good frequency-domain selectivity. More sophisticated filter pairs, with lengths >2, have been developed; these are designed to maintain the orthogonality and to impose additional conditions which ensure both time- and frequency-domain localization.

This set of tools can lead to fast algorithms for digital implementation of the central ideas in theoretical harmonic analysis.

A key point is that one can cascade the above operations. For example, if one can split a signal domain into two pieces, then one can split it into four pieces, by applying the same type of operation again to each piece. In this way, the dyadic structuring ideas that were so useful in harmonic analysis—dyadic partitioning of time-scale and time-frequency—correspond directly to dyadic structuring in the digital setting.

We give a few examples of this, stressing the role of combining elementary dyadic operations.

- *Fast Meyer Wavelet Transform:* The Meyer wavelet basis for $L^2(\boldsymbol{R})$ was originally defined by its frequency-domain properties, and so it is most natural to construct a digital variant using the Fourier domain. The forward transform algorithm goes as follows. Transform into the frequency domain. Apply smooth orthonormal windowing, breaking up the frequency domain into subbands of pairs of intervals of width 2^{j-1} samples, located symmetrically about zero frequency. Apply to each subband a Fourier analysis (actually either a sine or cosine transform) of length adapted to the length of the subband. The cost of applying this algorithm is dominated by the initial passage to the Frequency domain, which is order $O(N \log(N))$. The inverse transform systematically reverses these operations.

  The point of the fast algorithm is, of course, that one does not literally construct the basis functions, and one does not literally take the inner product of the digital signal with the basis function. This is all done implicitly. However, it is easy enough to use the algorithm to display the basis functions. When one does so, one sees that they are trigonometric polynomials which are periodic and effectively localized near the dyadic interval they should be associated with.

- *Mallat Algorithm:* Improving in several respects on the frequency-domain digital Meyer wavelet basis is a family of orthonormal wavelet bases based on time-domain filtering [69]. The central idea here is by now very well known: it involves taking the signal, applying subband partitioning with specially chosen digital highpass and lowpass filters, subsampling the two pieces by dyadic decimation, and then recursively applying the same procedure on the lowpass piece only. When the filters are appropriately specified, the result is an orthonormal transform on N-space.

  The resulting transform on digital signals takes only order N arithmetic operations, so it has a speed advantage over the fast Meyer wavelet transform, which requires order $N \log(N)$ operations. Although we do not describe it here, there is an important modification of the filtering operators at the ends of the sequence which allows the wavelet basis functions to adapt to the boundary of the signal, i.e., we avoid the periodization of the fast Meyer transform [13]. Finally, the wavelet basis functions are compactly supported, with basis element $\psi_{j,k}$ vanishing in the time domain outside an interval homothetic to $I_{j,k}$. This means that they have the correct structure to be

called a wavelet transform. The support property also means that a small number of wavelet coefficients at a given scale are affected by singularities, or to put it another way, the effect of a singularity is compressed into a small number of coefficients.

- *Fast Wilson Basis Transform* is a digital version of the Wilson basis; the fast transform works as follows. First, one applies a smooth orthonormal partitioning to break up the time domain into equal-length segments. (It is most convenient if the segments have dyadic lengths.) Then one applies Fourier analysis, in the form of a sine or cosine transform, to each segment. The whole algorithm is order $N \log(M)$, where M is the length of a segment. The implicitly defined basis functions look like windowed sinusoids with the same arrangement of sine and cosine terms as in the continuum Wilson basis.

We should now stress that for the correspondence between a theoretical harmonic analysis concept and a computational harmonic analysis tool, dyadic structuring operators should not be performed in a cavalier fashion. For example, if one is going to cascade subband partitioning operations many times, as in the Mallat algorithm, it is important that the underlying filters be rather specially chosen to be compatible with this repetitive cascade.

When this is done appropriately, one can arrive at digital implementations that are not vague analogies of the corresponding theoretical concepts, but can actually be viewed as "correct" digital realizations. As the reader expects by now, the mathematical expression that one has a "correct" realization is achieved by establishing a norm equivalence result. For example, if in the Mallat algorithm one cascades the subband partitioning operator using appropriate finite-length digital filters, one can construct a discrete wavelet transform for digital signals. This discrete transform is "correctly related" to a corresponding theoretical wavelet transform on the continuum because of a norm equivalence: if $f(t)$ is a function on the interval $[0,1]$, and if $(\tilde{\theta}_I)$ is a set of digital wavelet coefficients for the digitally sampled object $(f(n/N))$, then the appropriate Triebel and Besov norms of the digital wavelet coefficients behave quite precisely like the same norms of the corresponding initial segments of the theoretical wavelet coefficients (θ_I) of the continuum f. This is again a form of sampling theorem, showing that an ℓ^2 equality of norms (which follows simply from orthogonality) is accompanied by an equivalence in many other norms (which follows from much more delicate facts about the construction). A consequence, and of particular relevance for this paper, is that under simple assumptions, one can use the digital wavelet transform coefficients for nonlinear approximation and expect the same bounds on approximation measures to apply; *hence digital wavelet-based transform coding of function classes obeys the same types of estimates as theoretical wavelet-based transform coding*.

The "appropriate finite-length filters" referred to in the last paragraph are in fact arrived at by a delicate process of design. In [20], a collection of finite-length filters was constructed that gave orthonormal digital wavelet transforms. For the constructed transforms, if one views the N digital samples as occurring at points n/N for $0 \leq n < N$, and takes individual digital basis elements corresponding to the "same" location and scale at different dyadic N, appropriately normalized and interpolated to create functions of a continuous variable, one obtains a sequence of functions which tends to a limit. Moreover, the limit must be a translate and dilate of a single smooth function of compact support. In short, the digital wavelet transform is truly a digital realization of the theoretical wavelet transform. The norm equivalence statement of the previous paragraph is a way of mathematically completing this fundamental insight. In the last ten years, a large number of interesting constructions of "appropriate finite-length filters" have appeared, which we cannot summarize here. For more complete information on the properties of wavelet transforms and the associated filters, see for example, [100] and [23].

The success in developing ways to translate theoretical wavelet transforms and Gabor transforms into computationally effective methods naturally breeds the ambition to do the same in other cases. Consider the dyadic Heisenberg cells of Section XVIII-E, and the resulting concept of nonuniform tiling of the time-frequency plane. It turns out that for a large collection of nonuniform tilings, one can realize—in a computationally effective manner—a corresponding orthonormal basis [18]; see Fig. 9.

Here are two examples:

- *Cosine Packets:* Take a recursive dyadic partition of the time interval into dyadic segments. Apply smooth orthonormal partitioning to separate out the original signal into a collection of corresponding segments. Take appropriate finite Fourier transforms of each segment. The result is an orthogonal transform which is associated with a specific tiling of the time-frequency domain. That tiling has, for its projection on the time axis, simply the original partition of the time domain; over the whole time-frequency domain, it consists of columns whose widths are defined by intervals of the time-partition; within a column, the tiling simply involves congruent dyadic Heisenberg cells with specified width.
- *Wavelet Packets:* Take a recursive dyadic partition of the frequency interval into dyadic segments. Apply subband partitioning to separate out the original signal into a collection of corresponding frequency bands. The result is an orthogonal transform which costs at most $O(N)$ operations. It is associated with a specific tiling of the time-frequency domain. That tiling has, for its projection on the frequency axis, simply the original partition of the frequency domain; over the whole time-frequency domain, it consists of rows whose widths are defined by intervals of the frequency-partition; within a row, the tiling simply involves congruent dyadic Heisenberg cells with specified height.

Do these bases correspond to "correct" digital implementation of the theoretical partitioning? While they are orthogonal, we are not aware of results showing that they obey a wide range of other norm equivalences. It would be interesting to

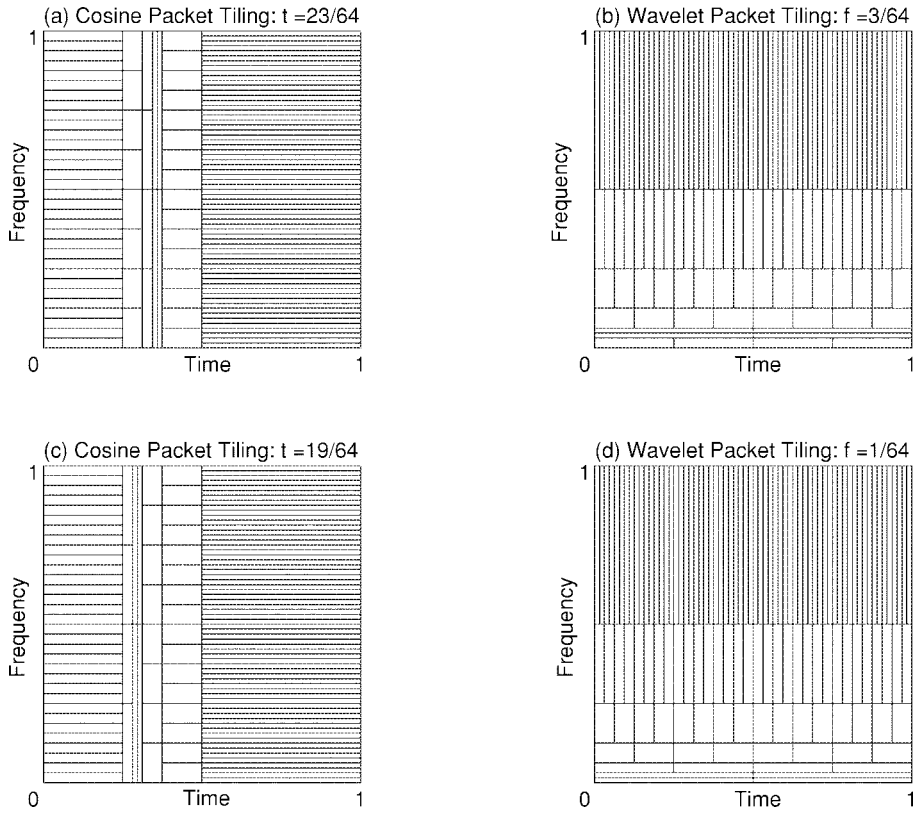

Fig. 9. Some time-frequency tilings corresponding to orthonormal bases.

know if they obey *sufficiently strong equivalences to imply that transform coding in an adaptively constructed basis can provide near-optimal coding* for an interesting class of objects. At the very least, results of this kind would require detailed assumptions, allowing the partitioning to be inhomogeneous in interesting ways and yet not very wild, and also supposing that the details of window lengths in the smooth orthonormal windowing or the filter choice in the subband partitioning are chosen appropriately.

The time-frequency tiling ideas raise interesting possibilities. In effect, the wavelet packet and cosine packet libraries create large libraries of orthogonal bases, all of which have fast algorithms. The Fourier and Wavelet bases are just two examples in this library; Gabor/Wilson-like bases provide other examples. These two collections have been studied by Coifman and Wickerhauser [17], who have shown that for certain "additive" objective functions the search through the library of all cosine packet or wavelet packet bases can be performed in $O(N \log(N))$ operations. This search is dependent on the function to be analyzed, so it is an instance of nonlinear approximation. In the case when this search is done for compression purposes, an operational rate-distortion-based version was presented in [80].

XX. Practical Coding

How do the ideas from harmonic analysis work on *real data*?

Actually, many of these ideas have been in use for some time in the practical data compression community, though discovered independently and studied under other names.

This is an example of a curious phenomenon commented on by Yves Meyer [73]: some of the really important ideas of harmonic analysis have been discovered, in some cognate or approximate form, in a wide range of practical settings, ranging from signal processing to mathematical physics, to computer-aided design. Often the basic tools are the same, but harmonic analysis imposes a different set of requirements on those tools.

To emphasize the relationship of the theory of this paper to practical coders, we briefly comment on some developments.

DCT coding of the type used in JPEG is based on a partitioning of the image domain into blocks, followed by DCT transform coding. The partitioning is done brutally, with the obvious drawback of DCT coding, which is the blocking effect at high compression ratios. Subband coding of images was proposed in the early 1980's [98], [102]. Instead of using rectangular windows as the DCT does, the subband approach effectively uses longer, smooth windows, and thus achieves a smooth reconstruction even at low bit rates. Note that early subband coding schemes were trying to approximate decorrelating transforms like the KLT, while avoiding some of the pitfalls of the DCT. Coding gains of filter banks were used as a performance measure. Underlying such investigations was a Gaussian model for signals, or a Markov random field (MRF) model for images.

Among possible subband coding schemes for images, a specific structure enjoyed particular popularity, namely, the scheme where the decomposition of the lower frequency band is iterated. This was due to several factors:

1) Its relationship with pyramid coding.

2) The fact that most of the energy remains concentrated in the lowpass version.
3) Its computational efficiency and simplicity.

Of course, this computational structure is equivalent to the discrete wavelet transform, and, as we have described, with appropriately designed digital filters, it can be related to the continuous wavelet series expansion.

Because of the concentrated efforts on subband coding schemes related to discrete wavelet transforms, interesting advances were achieved leading to improved image compression. The key insight derived from thinking in scale-location terms, and realizing that edges caused an interesting clustering effect across scales: the positions of "big" coefficients would remain localized in the various frequency bands, and could therefore be efficiently indexed using a linking across frequencies [68]. This idea was perfected by J. Shapiro into a data structure called an embedded zero tree [86]. Together with a successive approximation quantization, this coder achieves high-quality, successive approximation compression over a large range of bit rates, outperforming JPEG at any given bit rate by several decibels in signal-to-noise ratio. Many generalizations of Shapiro's algorithm have been proposed, and we will briefly outline the basic scheme to indicate its relation to nonlinear approximation in wavelet decompositions.

A key idea is that the significant wavelet coefficients are well-localized around points of discontinuity at small scales (or high frequency), see Fig. 10(a). This is unlike local cosine or DCT bases. Therefore, an edge of a given orientation will appear roughly as an edge in the respective orientation subband, and this at all scales. Conversely, smooth areas will be nulled out in passbands, since the wavelet has typically several zero moments. Therefore, a conditional entropy coder can take advantage of this "dependence across scales." In particular, the zero tree structure gathers regions of low energy across scales, by simply predicting that if a passband is zero at a given scale, its four children at the next finer scale (and similar orientation) are most likely to be zero as well (see Fig. 10(b)). This scheme can be iterated across scales. In some sense, it is a generalization of the end of block symbol used in DCT coding. (Note that such a prediction across scales would be fruitless in a Gaussian setting, because of independence of different bands.) Also, note that the actual values of the coefficients are *not* predicted, but only the "insignificance" or absence of energy; i.e., the idea is one of positional coding. Usually, the method is combined with successive approximation quantization, leading to an embedded bit stream, where successive approximation decoding is possible. An improved version, where larger blocks of samples are used together in a technique called set partitioning, lead to the SPIHT algorithm [82] with improved performance and lower computational complexity. Other variations include context modeling in the various bands [65]. Comparing these ideas with the present paper, we see a great commonality of approach to the use of trees for positional coding of the "big" wavelet coefficients as we have discussed in Section XVII-B.

At a more abstract level, such an approach to wavelet image coding consists in picking a subset of the largest coefficients of the wavelet transform, and making sure that the cost of addressing these largest coefficients is kept down by a smart conditional entropy code. That is, the localization property of the wavelet transform is used to perform efficient addressing of singularities, while the polynomial reproduction property of scaling functions allows a compact representation of smooth surfaces.

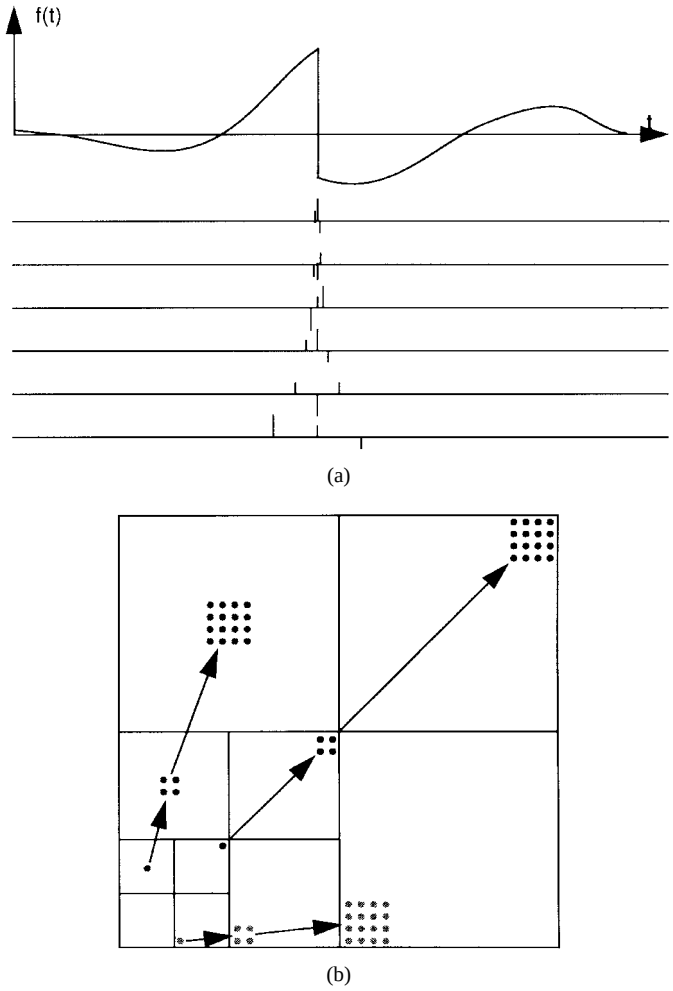

Fig. 10. Localization of the wavelet transform. (a) One-dimensional signal with discontinuity. (b) Two-dimensional signal and linking of scale-space points used in wavelet image coding (EZW and SPIHT).

How about the performance of this coder and its related cousins? In Fig. 3, we see that a substantial improvement is achieved over JPEG (the actual coder is from [82]). However, the basic behavior is the same, that is, at fine quantization, we find again the typical $D(R) \sim 2^{-2R}$ slope predicted by classical high rate-distortion theory. Instead, we would hope for a decay related to the smoothness class. At low bit rate, a recent analysis by Mallat and Falzon [70] shows $D(R) \sim R^{-1}$ by modeling the nonlinear approximation features of such a wavelet coder.

The next generation image coding standard, JPEG-2000, is considering schemes similar to what was outlined above. That is, it has been proposed to exploit properties of the wavelet transform and of the structure of wavelet coefficients across scales, together with state-of-the-art quantization (using, for example, trellis-coded quantizers) and adaptive entropy coders.

In short, there are a variety of interesting parallels between practical coding work and the work in harmonic analysis. We are aware of other areas where interesting comparisons can be made, for example, in speech coding, but omit a full comparison of literatures for reasons of space.

XXI. Summary and Prognosis

In composing this survey, we have been inspired by an attractive mixture of ideas. To help the reader, we find it helpful to summarize these ideas, and to make them memorable by associating them with prominent figures of this century.

A. Pure Mathematics Disdaining Usefulness

G. H. Hardy is a good example for this position. In *A Mathematician's Apology*, he gave a famous evaluation of his life's work:

> "I have never done anything 'useful.' No discovery of mine has made, or is likely to make, directly or indirectly, for good or for ill, the least difference to the world."

In the same essay, he argued strenuously against the proposition that pure mathematics could ever have "utility."

The irony is, of course, that, as discussed above, the purely mathematical impulse to understand Hardy H^p spaces gave rise to the first orthonormal basis of smooth wavelets.

In this century, harmonic analysis has followed its own agenda of research, involving among other things the understanding of equivalent norms in functional spaces. It has developed its own structures and techniques for addressing problems of its own devising; and if one studies the literature of harmonic analysis one is struck by the way the problems—e.g., finding equivalent norms for H^p spaces—seem unrelated to any practical needs in science or engineering—e.g., H^p spaces do not consist of "objects" which could be understood as modeling a collection of "naturally occurring objects" of "practical interest." Nor should this be surprising; David Hilbert once said that "the essence of mathematics is its freedom" and we suspect he means "freedom from any demands of the outside world to be other than what its own internal logic demands."

Despite the freedom and in some sense "unreality" of its orientation, the structures and techniques that harmonic analysis has developed in pursuing its agenda have turned out to be significant for practical purposes—in particular, as mentioned above, discrete wavelet transforms are being considered for inclusion in future standards like JPEG-2000.

In this paper we have attempted to "explain" how an abstractly oriented endeavor could end up interacting with practical developments in this way. In a sense, harmonic analysts, by carrying out their discipline's internal agenda, discovered a way to "diagonalize" the structure of certain functional classes outside of the L^2 cases where this concept arose. This "diagonalization" carries a data compression interpretation, and leads to algorithmic ideas for dealing with other kinds of data than traditional Gaussian data compression theory allows.

B. Stochastic versus Deterministic Viewpoints

Even if one accepts that pure mathematics can have unexpected outcomes of relevance to practical problems, it may seem unusual that analysis *per se*—which concerns deterministic objects—could be connected with the compression of real data, which concerns random objects. To symbolize this possibility, we make recourse to one of the great mathematicians of this century, A. N. Kolmogorov. V. M. Tikhomirov, in an appreciation of Kolmogorov, said [93]

> ... "our vast mathematical world" is itself divided into two parts, as into two kingdoms. Deterministic phenomena are investigated in one part, and random phenomena in the other.
>
> To Kolmogorov fell the lot of being a trailblazer in both kingdoms, a discoverer in their many unexplored regions ... he put forth a grandiose programme for a simultaneous and parallel study of the complexity of deterministic phenomena and the uncertainty of random phenomena, and the experience of practically all his creative biography was concentrated in this programme.
>
> From the beginning of this programme, the illusory nature of setting limits between the world of order and the world of chance revealed itself.

The scholarly record makes the same point. In his paper at the 1956 IRE Conference on Information Theory, held at MIT, Kolmogorov [60] published the "reverse water-filling formula" giving $R(D)$ for Gaussian processes (2.5), (2.6), which as we have seen is the formal basis for transform coding. He also described work with Gel'fand and Yaglom rigorously extending the sense of the mutual information formula (2.4) from finite-dimensional vectors X and Y to functional data. These indicate that the functional viewpoint was very much on his mind in connection with the Shannon theory. In that same paper he also chose to mention the ϵ-entropy concept which he had recently defined [61], and he chose a notation which allowed him to make the point that ϵ-entropy is formally similar to Shannon's $R(D)$. One gets the sense that Kolmogorov thought the two theories might be closely related at some level, even though one concerns deterministic objects and the other concerns random objects; see also his return to this theme in more depth in the appendix in the monograph [62].

C. Analysts Return to Concrete Arguments

Shannon's key discoveries in lossy data compression are summarized in (2.2)–(2.4). These are highly abstract and in fact can be proved in an abstract setting; compare the abstract alphabet source coding theorem and converse in Berger (1971). In this sense, Shannon was a man of his time. During the 1930's–1950's, abstraction in mathematical science was in full bloom; it was the era that produced the formalist mathematical school of Bourbaki and fields like "abstract harmonic analysis."

Kolmogorov's work in lossy data compression also was a product of this era; the concept of ϵ-entropy is, to many newcomers' tastes, abstraction itself.

Eventually the pendulum swung back, in a return to more concrete arguments and constructions, as illustrated by the work of the Swedish mathematician Lennart Carleson. In a very distinguished research career spanning the entire period from 1950 to the present, Carleson obtained definitive results of lasting value in harmonic analysis, the best known of which is the almost everywhere convergence of Fourier series, an issue essentially open since the 19th century, and requiring a proof which has been described by P. Jones as "one of the most complicated to be found in modern analysis" [57]. Throughout his career, Carleson created new concepts and techniques as part of resolving very hard problems; including a variety of dyadic decomposition ideas and the concept of Carleson measures. It is interesting to read Carleson's own words [11].

> There was a period, in the 1940's and 1950's, when classical analysis was considered dead and the hope for the future of analysis was considered to be in the abstract branches, specializing in generalization. As is now apparent, the death of classical analysis was greatly exaggerated and during the 1960's and 1970's the field has been one of the most successful in all of mathematics. ... the reasons for this ... [include] ... the realization that in many problems complications cannot be avoided, and that intricate combinatorial arguments rather than polished theories are in the center.

Carleson "grew up" as a young mathematician in the early 1950's, and so it is natural that he would react against the prevailing belief system at the time of his intellectual formation. That system placed great weight on abstraction and generality; Carleson's work, in contrast, placed heavy emphasis on creating useful tools for certain problems which by the standards of abstract analysts of the day, were decidedly concrete.

Carleson can be taken as symbolic of the position that a concrete problem, though limited in scope, can be very fertile.

D. The Future?

So far, we have used important personalities to symbolize progress to date. What about the future?

One of the themes of this paper is that harmonic analysts, while knowingly working on hard problems in analysis, and discovering tools to prove fundamental results, have actually been developing tools with a broad range of applications, including data compression. Among harmonic analysts, this position is championed by R. R. Coifman. His early work included the development of atomic decompositions for H^p spaces, $p \leq 1$. Today, his focus is in another direction entirely, as he develops ways to accelerate fundamental mathematical algorithms and to implement new types of image compression. This leads him to reinterpret standard harmonic analysis results in a different light.

For example, consider his attitude toward a milestone of classical analysis: L. Carleson's proof of the almost-everywhere convergence of Fourier series, which is generally thought of as a beautiful and extremely complex pure analysis argument. But apparently Coifman sees here a serious effort to understand the underlying "combinatorics" of time and frequency bases, a "combinatorics" potentially also useful (say) for "time-frequency-based signal compression."

In another direction, consider the work of P. Jones [56] who established a beautiful result showing that one can approximately measure the length of the traveling salesman tour of a set of points in the plane by a kind of nonlinear Littlewood–Paley analysis. (This has far-reaching extensions by David and Semmes [24].) Others may see here the beginnings of a theory of quantitative geometric measure theory. Coifman apparently sees a serious effort to understand the underlying combinatorics of curves in the plane (and in David–Semmes, hypersurfaces in higher dimensional spaces), a "combinatorics" which is potentially also useful (say) for compressing two- and higher dimensional data containing curvilinear structure.

The position underlying these interpretations is exactly the opposite of Hardy: Hardy believed that his research would not be "useful" because he did not *intend* it to be; yet, it turns out that research in harmonic analysis has been and may well continue to be "useful" even when researchers, like Hardy, have no conscious desire to be useful.

How far can the connection of Harmonic Analysis and Data Compression go? We are sure it will be fun to find out.

ACKNOWLEDGMENT

The authors would like to thank R. R. Coifman for exposing us to many iconoclastic statements which were provocative, initially mysterious, and ultimately very fruitful. They would also like to thank the editors of this special issue. Sergio Verdú had the idea for assembling this team for this project; both he and Stephen McLaughlin displayed great patience with us in the stages near completion.

D. L. Donoho would like to thank Miriam Donoho for vigorously supporting the effort to make this paper appear. He would also like to thank Xiaoming Huo for extensive effort in preparing several figures for this paper.

M. Vetterli would like to thank Robert M. Gray, Vivek Goyal, Jelena Kovačević, Jérôme Lebrun, Claudio Weidmann, and Bin Yu for interactions and help.

I. Daubechies would like to thank Nela Rybowicz for her extraordinary helpfulness in correcting the page proofs.

REFERENCES

[1] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput.*, vol. C-23, pp. 88–93, Jan. 1974.

[2] E. W. Aslaksen and J. R. Klauder, "Unitary representations of the affine group," *J. Math. Phys.*, vol. 9, pp. 206–211, 1968.

[3] P. Auscher, G. Weiss, and G. Wickerhauser, "Local sine and cosine bases of Coifman and Meyer and the construction of smooth wavelets," in *Wavelets: A Tutorial in Theory and Applications*, C. K. Chui, Ed. New York: Academic, 1992.

[4] H. B. Barlow, "Possible principles underlying the transformation of sensory messages," in *Sensory Communication*, W. A. Rosenbluth, Ed. Cambrige, MA: MIT Press, 1961, pp. 217–234.

[5] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice Hall, 1971.

[6] L. Birgé, "Approximation dans les espaces métriques et théorie de l'estimation," (in French), ("Approximation in metric spaces and the theory of estimation"), *Z. Wahrsch. Verw. Gebiete*, vol. 65, no. 2, pp. 181–237, 1983.

[7] L. Birgé and P. Massart, "An adaptive compression algorithm in Besov spaces," *Constr. Approx.*, to be published.

[8] M. S. Birman and M. Z. Solomjak, "Piecewise-polynomial approximations of functions of the classes W_p^α," *Mat. Sbornik*, vol. 73, pp. 295–317, 1967.

[9] A. P. Calderón, "Intermediate spaces and interpolation, the complex method," *Studia Math.*, vol. 24, pp. 113–190, 1964.

[10] B. Carl and I. Stephani, *Entropy, Compactness, and the Approximation of Operators*. Cambridge, U.K.: Cambridge Univ. Press, 1990.

[11] L. Carleson, "The work of Charles Fefferman," in *Proc. Int. Congr. Mathematics* (Helsinki, Finland, 1978). Helsinki, Finland: Acad. Sci. Fennica, 1980, pp. 53–56.

[12] A. Cohen and J. P. D'Ales, "Non-linear approximation of random functions," *SIAM J. Appl. Math.*, vol. 57, pp. 518–540, 1997.

[13] A. Cohen, I. Daubechies, and P. Vial, "Orthonormal wavelets for the interval," *Appl. Comput. Harmonic Anal.*, vol. 1, no. 1, 1993.

[14] A. Cohen, I. Daubechies, O. Guleryuz, and M. Orchard, "On the importance of combining wavelet-based nonlinear approximation in coding strategies," unpublished manuscript, 1997.

[15] A. Cohen, W. Dahmen, I. Daubechies, and R. A. DeVore, "Tree, Approximation and Encoding," preprint, 1998..

[16] R. R. Coifman, "A real-variable characterization of H^p," *Studia Math.*, vol. 51, pp. 269–274, 1974.

[17] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," vol. 38, no. 2, pp. 713–718, Mar. 1992.

[18] R. R. Coifman and Y. Meyer, "Remarques sur l'analyze de Fourier à fenêtre, *C. R. Acad Sci. Paris*, vol. 312, pp. 259–261, 1991.

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

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

[21] ———, "The wavelet transform, time-frequency localization, and signal analysis," *IEEE Trans. Inform. Theory*, vol. 36, pp. 961–1005, 1990.

[22] I. Daubechies, S. Jaffard, and J. L. Journé, "A simple Wilson orthonormal basis with exponential decay," *SIAM J. Math. Anal.*, vol. 24, pp. 520–527, 1990.

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

[24] G. David and S. Semmes, *Analysis of and on Uniformly Rectifiable Sets.* (Amer. Math. Soc., Mathematical Surveys and Monographs, vol. 38), 1993.

[25] K. DeLeeuw, J. P. Kahane, and Y. Katznelson, "Sur les coefficients de Fourier des fonctions continues," *C. R. Acad. Sci. Paris*, vol. 285, pp. 1001–1003, 1978.

[26] R. A. DeVore, "Nonlinear approximation," *Acta Numer.*, vol. 7, pp. 51–150, 1998.

[27] R. A. DeVore, B. Jawerth, and B. Lucier, "Image compression through wavelet transform coding," *IEEE Trans. Inform. Theory*, vol. 38, pp. 719–746, 1992.

[28] R. DeVore and X. M. Yu, "Degree of adaptive approximation," *Math. Comput.*, vol. 55, pp. 625–635, 1990.

[29] R. DeVore, B. Jawerth, and V. A. Popov, "Compression of wavelet decompositions," *Amer. J. Math.*, vol. 114, pp. 737–785, 1992.

[30] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*. New York: Springer, 1993.

[31] D. L. Donoho, "Unconditional bases are optimal bases for data compression and statistical estimation," *Appl. Comput. Harmonic Anal.*, vol. 1, no. 1. pp. 100–105, 1993.

[32] ———, "Unconditional bases and bit-level compression," *Appl. Comput. Harmonic Anal.*, vol. 3, no. 4, pp. 388–392, 1996.

[33] ———, "CART and best-ortho-basis: A connection," *Ann. Statist.*, vol. 25, no. 5, pp. 1870–1911, 1997.

[34] ———, "Counting bits with Kolmogorov and Shannon," manuscript, 1998.

[35] R. M. Dudley, "The sizes of compact subsets of Hilbert spaces and continuity of Gaussian processes," *J. Funct. Anal.*, vol. 1, pp. 290–330, 1967.

[36] D. E. Edmunds and H. Triebel, *Function Spaces, Entropy Numbers, and Differential Operators*. Cambridge, U.K.: Cambridge Univ. Press, 1996.

[37] D. J. Field, "The analysis of natural images and the response properties of cortical cells," *J. Opt. Soc. Amer.*, 1987.

[38] ———, "What is the goal of sensory coding?" *Neural Comput.*, vol. 6, no. 4, pp. 559–601, 1994.

[39] G. Folland, *Harmonic Analysis in Phase Space*. Princeton, NJ: Princeton Univ. Press, 1989.

[40] H. G. Feichtinger, "Atomic characterizations of modulation spaces through Gabor-type representations," *Rocky Mount. J. Math.*, vol. 19, pp. 113–126, 1989.

[41] C. Fefferman, "The uncertainty principle," *Bull. Amer. Math. Soc.*, vol. 9, pp. 129–206, 1983.

[42] M. Frazier and B. Jawerth, "The ϕ-transform and applications to distribution spaces," in *Function Spaces and Applications* (Lecture Notes in Mathematics, vol. 1302), M. Cwikel, *et al.*, Eds. Berlin, Germany: Springer-Verlag, 1988, pp. 223–246.

[43] M. Frazier, B. Jawerth, and G. Weiss, *Littlewood–Paley Theory and the Study of Function Spaces* (CBMS Reg. Conf. Ser. in Mathematics, no. 79). Amer. Math. Soc., 1991.

[44] D. Gabor, "Theory of communication," *J. Inst. Elect. Eng.*, vol. 93, pp. 429–457, 1946.

[45] J. Garnett, *Bounded Analytic Functions*. New York: Academic, 1981.

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

[47] R. Godement, "Sur les relations d'orthogonalité de V. Bargmann," *C. R. Acad. Sci. Paris*, vol. 255, pp. 521–523, 657–659, 1947.

[48] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inform. Theory*, this issue, pp. 2325–2383.

[49] V. Goyal and M. Vetterli "Computation-distortion characteristics of block transform coding," in *ICASSP-97* (Munich, Germay, Apr. 1997), vol. 4, pp. 2729–2732.

[50] K. Gröchenig and S. Samarah, "Nonlinear approximation with local Fourier bases," unpublished manuscript, 1998.

[51] K. Gröchenig and D. Walnut, "Wilson bases are unconditional bases for modulation spaces," unpublished manuscript, 1998.

[52] A. Grossmann and J. Morlet, "Decomposition of Hardy functions into square-integrable wavelets of constant shape," *SIAM J. Appl. Math.*, vol. 15, pp. 723–736, 1984.

[53] G. H. Hardy, *A Mathematician's Apology*. Cambridge, U.K.: Cambridge Univ. Press, 1940.

[54] R. Howe, "On the role of the Heisenberg group in hramonic analysis." *Bull Amer. Math. Soc.*, vol. 3, pp. 821–843, 1980.

[55] J. J. Y. Huang and P. M. Schultheiss, "Block quantization of correlated Gaussian random variables," *IEEE Trans. Commun.*, vol. CUM-11, pp. 289–296, Sept. 1963.

[56] P. Jones, "Rectifiable sets and the travelling salesman problem," *Inventiones Mathematicae*, vol. 102, pp. 1–15, 1990.

[57] ———, "Lennart Carleson's work in analysis," in *Festschrift in Honor of Lennart Carleson and Yngve Domar*, (*Acta Universitatis Upsaliensis*, vol. 58). Stockholm, Sweden: Almquist and Wiksell Int., 1994.

[58] J. P. Kahane and P. G. Lemarié-Rieusset, *Fourier Series and Wavelets*. Luxemburg: Gordon and Breach, 1995.

[59] J. Klauder and B. Skagerstam, *Coherent States, Applications in Physics and Mathematical Physics*. Singapore: World Scientific, 1985.

[60] A. N. Kolmogorov, "On the Shannon theory of information transmission in the case of continuous signals," *IRE Trans. Inform. Theory*, vol. IT-2, pp. 102–108, 1956.

[61] ———, "Some fundamental problems in the approximate and exact representation of functions of one or several variables," in *Proc III. Math Congress USSR*, vol. 2. Moscow, USSR: MCU Press, 1956, pp. 28–29. Reprinted in Komogorov's *Selected Works*, vol. I.

[62] A. N. Kolmogorov and V. M. Tikhomirov, ϵ-entropy and ϵ-capacity. *Usp. Mat. Nauk*, vol. 14, pp. 3–86, 1959. (English transl. *Amer. Math. Soc. Transl.*, ser. 2, vol 17, pp. 277–364.)

[63] H. P. Kramer and M. V. Mathews, "A linear coding for transmitting a set of correlated signals," *IRE Trans. Inform. Theory*, vol. IT-23, pp. 41–46, Sept. 1956.

[64] P. G. Lemarié and Y. Meyer, "Ondelettes et Bases Hilbertiennes," *Revista Mat. Iberomamericana*, vol. 2, pp. 1–18, 1986.

[65] S. LoPresto, K. Ramchandran, and M. T. Orchard, "Image coding based on mixture modeling of wavelet coefficients and a fast estimation-quantization framework," in *Data Compression Conf. '97* (Snowbird, UT, 1997), pp. 221–230.

[66] G. G. Lorentz, "Metric entropy and approximation," *Bull. Amer. Math. Soc.*, vol. 72, pp. 903–937, Nov. 1966.

[67] L. Le Cam, "Convergence of estimates under dimensionality restrictions," *Ann. Statist.*, vol. 1, pp. 38–53, 1973.

[68] A. S. Lewis and G. Knowles, "Image compression using the 2-D wavelet transform," *IEEE Trans. Image Processing*, vol. 1, pp. 244–250, Apr. 1992.

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

[70] S. Mallat and F. Falzon, "Analysis of low bit rate image transform coding," *IEEE Trans. Signal Processing*, vol. 46, pp. 1027–1042, Apr. 1998.

[71] H. S. Malvar and D. H. Staelin, "The LOT: Transform coding without blocking effects," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 4, pp. 553–559, 1989.

[72] H. S. Malvar, "Extended lapped transforms: Properties, applications, and fast algorithms," *IEEE Trans. Signal Processing*, vol. 40, pp. 2703–2714, 1992.
[73] Y. Meyer, "Review of 'An Introduction to Wavelets' and 'Ten Lectures on Wavelets'," *Bull. Amer. Math. Soc.*, vol. 28, pp. 350–359, 1993.
[74] ———, *Ondelettes et Operateurs*. Paris, France: Hermann, 1990.
[75] ———, "Wavelets and applications" (Lecture at CIRM Luminy meeting, Luminy, France, Mar. 1992).
[76] F. Mintzer, "Filters for distortion-free two-band multirate filter banks," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, pp. 626–630, June 1985.
[77] A. Pinkus, *n-Widths in Approximation Theory*. New York: Springer, 1983.
[78] G. Pisier, *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge, U.K.: Cambridge Univ.Press, 1989.
[79] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*. New York: Springer, 1997.
[80] 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.
[81] D. L. Ruderman, "Origins of scaling in natural images" *VISION RES.*, vol. 37, no. 23, pp. 3385–3398, Dec. 1997.
[82] A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierachical trees," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, pp. 243–250, June 1996.
[83] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.
[84] ———, "Communication in the presence of noise," *Proc. IRE*, vol. 37, pp. 10–21, 1949.
[85] ———, "The bandwagon" (1956), in *Claude Elwood Shannon: Collected Papers*, N. J. A. Sloane and A. D. Wyner, Eds. Piscataway, NJ: IEEE Press, 1993.
[86] J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Processing*, vol. 41, pp. 3445–3462, Dec. 1993.
[87] E. Simoncelli, "Statistical models for images: Compression, restoration and synthesis,"presented at the IEEE 31st Asilomar Conf. Signals, Systems, and Computers, Pacific Grove, CA, 1997.
[88] M. J. T. Smith and T. P. Barnwell, III, "Exact reconstruction for tree-structured subband coders," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 35, pp. 431–441, 1986.
[89] E. Stein, *Singular Integrals and Differentiability Properties of Functions*. Princeton, NJ: Princeton Univ. Press, 1970.
[90] ———, *Harmonic Analysis: Real Variable Methods, Orthogonality, and Oscillatory Integrals*. Princeton, NJ: Princeton Univ. Press, 1993.
[91] J. O. Stromberg, *Festschrift in Honor of Antoni Zygmund*. Monterey, CA: Wadsworth, 1982.
[92] ———, "Maximal functions, Hardy Spaces, BMO, and Wavelets, from my point of view," in *Festschrift in Honor of Lennart Carleson and Yngve Domar*, (*Acta Universitatis Upsaliensis* vol. 58). Stockholm, Sweden: Almquist and Wiksell Int., 1994.
[93] V. M. Tikhomirov, "Widths and entropy," *Usp. Mat. Nauk*, vol. 38, pp. 91–99, 1983 (in Russian). English transl. in *Russian Math Surveys*, vol. 38, pp. 101–111.
[94] ———, "Commentary: ϵ-entropy and ϵ-capacity," in *A. N. Kolmogorov: Selected Works. III. Information Theory and Theory of Algorithms*, A. N. Shiryaev, Ed. Boston, MA: Kluwer, 1992.
[95] B. Torrésani, "Time-frequency representations: Wavelet packets and optimal decomposition," *Ann. l'Institut Henri Poincare* (*Physique Théorique*), vol. 56, no. 2, pp. 215–34, 1992.
[96] H. Triebel, *Theory of Function Spaces*. Basel, Switzerland: Birkhauser, 1983.
[97] P. P. Vaidyanathan, "Quadrature mirror filter banks, M-band extensions and perfect reconstruction techniques," *IEEE ASSP Mag.*, vol. 4, pp. 4–20, July 1987.
[98] M. Vetterli, "Multi-dimensional subband coding: Some theory and algorithms," *Signal Processing*, vol. 6, no. 2, pp. 97–112, Apr. 1984.
[99] ———, "Filter banks allowing perfect reconstruction," *Signal Processing*, vol. 10, no. 3, pp. 219–244, Apr. 1986.
[100] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding*. Englewood Cliffs, NJ: Prentice-Hall, 1995.
[101] P. H. Westerink, J. Biemond, D. E. Boekee, and J. W. Woods, "Subband coding of images using vector quantization," *IEEE Trans. Commun.*, vol. 36, pp. 713–719, June 1988.
[102] J. W. Woods and S. D. O'Neil, "Sub-band coding of images," *IEEE Trans. Acoust., Speech Siganl Processing*, vol. 34, pp. 1278–1288, Oct. 1986.
[103] A. Zygmund, *Trigonometric Series*, vols. I & II. Cambridge, U.K.: Cambridge Univ. Press, 1959.