On the Importance of Combining Wavelet-Based Nonlinear Approximation With Coding Strategies

Albert Cohen, *Member, IEEE*, Ingrid Daubechies, *Fellow, IEEE*, Onur G. Guleryuz, *Member, IEEE*, and Michael T. Orchard, *Fellow, IEEE*

***Abstract*—This paper provides a mathematical analysis of transform compression in its relationship to linear and nonlinear approximation theory. Contrasting linear and nonlinear approximation spaces, we show that there are interesting classes of functions/random processes which are much more compactly represented by wavelet-based nonlinear approximation. These classes include locally smooth signals that have singularities, and provide a model for many signals encountered in practice, in particular for images. However, we also show that nonlinear approximation results do not always translate to efficient compression strategies in a rate-distortion sense. Based on this observation, we construct compression techniques and formulate the family of functions/stochastic processes for which they provide efficient descriptions in a rate-distortion sense. We show that this family invariably leads to Besov spaces, yielding a natural relationship among Besov smoothness, linear/nonlinear approximation order, and compression performance in a rate-distortion sense. The designed compression techniques show similarities to modern high-performance transform codecs, allowing us to establish relevant rate-distortion estimates and identify performance limits.**

***Index Terms*—Besov spaces, linear approximation, nonlinear approximation, rate-distortion, transform coding, wavelets.**

I. Introduction

In theoretical models for the mathematical study of compression, signals and particularly images are often viewed as realizations of an (unknown) stochastic process. The corresponding Karhunen–Loève (KL) basis, as the orthonormal basis that optimally decorrelates this process, i.e., the basis $(\varphi_n)_{n\in\mathbb{N}}$ that minimizes

$$E\left(\left\|f-\sum_{n=1}^{N}\langle f,\,\varphi_n\rangle\varphi_n\right\|^2\right)$$

for every N, is then viewed as the best possible basis to compress the signals or images. In practice, determining this KL basis exactly may be cumbersome and computationally intensive, suggesting the use of a basis that is easier to work with and that is still "close" to the KL basis, in the sense that it also decorrelates well (although not optimally). This has been argued as a justification both for discrete cosine transform (DCT) methods and for wavelet transforms.

Although the usefulness of KL bases is well documented and beyond dispute in many applications, there has been a growing realization that optimizing decorrelation for the stochastic process may not be the final or even the most important point in signal compression. In the terms of mathematical approximation theory, this corresponds to a shift from linear approximation to nonlinear approximation.

Let us try to describe in a nutshell the difference between linear approximation and nonlinear approximation. In linear approximation theory, given an orthonormal basis of functions $\varphi_1, \varphi_2, \ldots$, one seeks to estimate, as a function of N, the truncation error

$$\varepsilon_N(f)=\left\|f-\sum_{n=1}^{N}\langle f,\,\varphi_n\rangle\varphi_n\right\|^2 \tag{1}$$

for a deterministic function f. The behavior of $\varepsilon_N(f)$ as N increases gives us information about f and *vice versa*. For instance, if the φ_n are either the Fourier basis on $[0,\,1]$ or a wavelet basis (with its logical ordering), and if $\|\,\|$ in (1) is taken to be the $\mathcal{L}^2$-norm for functions in the unit interval, $\|g\|=(\int_0^1|g(t)|^2\,dt)^{1/2}$, then the decay of $\varepsilon_N(f)$ characterizes the smoothness of f in an $\mathcal{L}^2$-sense: $\varepsilon_N(f)\le CN^{-2k-\eta}$ for some $\eta>0$ implies that $f\in W^{k,2}$, i.e., that f lies in a Sobolev space and its first k derivatives are all in $\mathcal{L}^2$, which itself implies $\varepsilon_N(f)\le CN^{-2k}$ (more details are given in Section II). The error $\varepsilon_N(f)$ can be rewritten as

$$\varepsilon_N(f)=\mathrm{dist}_{\mathcal{L}^2}(f,\,\mathcal{S}_N)^2$$

where $\mathcal{S}_N$ is the linear vector space spanned by the first N basis functions

$$\mathcal{S}_N=\mathrm{Span}(\varphi_1,\,\ldots,\,\varphi_N)=\left\{\sum_{n=1}^{N}c_n\varphi_n;\,c_n\in\mathbb{C}\right\}$$

and $\mathbb{C}$ denotes the set of complex numbers. The KL basis for a stochastic process fits within this linear approximation theory framework. For a stochastic process f, it is the basis for which $E(\mathrm{dist}_{\mathcal{L}^2}(f,\,\mathcal{S}_N)^2)$ is minimized, for every N.

Nonlinear approximation of a function f, with the same orthonormal basis $\varphi_1, \varphi_2, \ldots$ as before, seeks to estimate, as N increases, the decay of the distance between f and the best possible approximation to f by a linear combination of φ_n that uses

Manuscript received January 7, 2000; revised January 24, 2002. The work of O. G. Guleryuz and M. T. Orchard was supported by the Army Research Office under Grant DAA-HO496-1-0227.

A. Cohen is with the Laboratoire d'Analyze Numerique, Universite Pierre et Marie Curie, 5252 Paris Cedex 05, France (e-mail: cohen@ann.jussieu.fr).

I. Daubechies is with the Department of Mathematics, Princeton University, Engineering Quadrangle, Princeton, NJ 08544 USA (e-mail: ingrid@math.princeton.edu).

O. G. Guleryuz is with the Epson Palo Alto Laboratory, Palo Alto, CA 94304 USA (e-mail: oguleryuz@erd.epson.com).

M. T. Orchard is with the Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005 USA (e-mail: orchard@ece.rice.edu).

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

Publisher Item Identifier S 0018-9448(02)05170-2.

at most N terms (as opposed to the best linear combination with the *first* N basis functions, as before). That is, we now have

$$\varepsilon_N(f) = \mathrm{dist}_{\mathcal{L}^2}(f, \mathcal{T}_N)^2$$
$$\mathcal{T}_N = \left\{ \sum_n c_n \varphi_n;\ c_n \in \mathbb{C},\ \#\{n | c_n \neq 0\} \leq N \right\}.$$

The set $\mathcal{T}_N$ is no longer a linear space (since the sum of two arbitrary elements of $\mathcal{T}_N$ generally uses more than N basis functions), hence the name *nonlinear* approximation. Nonlinear approximation in a basis (as well as other types of nonlinear approximation, e.g., by rational functions or by free knot splines) has been studied in various contexts, see [8] for a substantial review. Using nonlinear approximation one typically obtains approximation estimates that are different from those obtained using linear approximation with respect to the same basis. Or turning this upside down, in nonlinear approximation a given decay behavior of the truncation error as $N \to \infty$ characterizes different behavior of f than it would in linear approximation. Another manifestation of this difference is that for stochastic processes, the KL basis need not be the basis that minimizes the nonlinear approximation error.

The difference between linear and nonlinear approximation was illustrated in [2] where a one-dimensional (1-D) model of piecewise-smooth processes, inspired by images, was analyzed. For this toy model, it was shown that the expected nonlinear approximation mean-square error using a wavelet basis is asymptotically superior (with a decay *at least* as fast as CN^{-2} for $N \to \infty$) to the best expected linear approximation error, even though wavelets are not the KL basis for the stochastic process. The linear approximation error, with respect to the KL basis, is *no better than* CN^{-1} for $N \to \infty$.

The success of wavelet bases in nonlinear approximation, as first analyzed in [9] and illustrated by the piecewise-smooth toy model in [2], was interpreted by mathematicians to be the "true" reason of the usefulness of wavelets in signal compression, rather than their potential for decorrelation. Yet nonlinear approximation estimates are still a long way from a mathematical analysis that would be directly related to compression issues. The compression practice is *not*, as suggested by nonlinear approximation, to "compress" all the information into N coefficients, discarding all the other information, and checking how well one does. A model closer to practice would estimate the error or distortion on the condition that all the information (truncated/quantized coefficients as well as the choice of which indexes to retain) has to fit within a certain bit budget. Such "operational rate-distortion" estimates are closer to the practice of coding for compression, and therefore more convincing from the point of view of engineers. On the other hand, coding techniques for compression have become increasingly sophisticated in the last few years, steadily obtaining better results. Some of the improved strategies for wavelet or subband image coders, such as [22] or [21], were inspired by, or can be heuristically explained by mathematical arguments. Yet there exist no mathematical estimates for these coding strategies of the same level of detail and depth as for the coding-wise more naive nonlinear approximation theory. The goal of this paper is to provide such a mathematical analysis.

This paper is structured as follows. First of all, in Section II, we recall several results from mathematical approximation theory on linear versus nonlinear approximation, both for functions and stochastic processes "living" in various Besov spaces. We also review in this section the results proved in [2] which contrast, for a particular 1-D model of a piecewise-smooth process (called PSM in the remainder of this paper), the best possible linear approximation with a qualitatively much better nonlinear approximation. The point we want to make here is not that the PSM model should be taken completely seriously for the modeling of images. But, since one can *prove* for this simplified 1-D "caricature" model of image rows (or columns) that nonlinear approximation works better than KL transforms, the same superiority of nonlinear approximation over KL transforms can reasonably be expected in less simplified mathematical representations, closer to real images.

As pointed out above, nonlinear approximation theorems do not give coding algorithms. In order to make a transition from approximation theory to results in a rate-distortion sense, we must discuss bit allocation, and see how the approximation bounds fare. This is the subject of Section III. In the case of nonlinear approximation with a wavelet basis, the wavelet coefficients are effectively "re-ordered" according to how significant they are in reducing the approximation error. Hence, for a given number of retained wavelet coefficients in nonlinear approximation, we also need to evaluate the number of bits necessary to encode the "addressing" of these coefficients, i.e., the number of bits necessary to convey which coefficients have been retained. Thus, for nonlinear approximation the number of required bits comes in two parts: bits for the quantized coefficients and bits for the addressing information. We show in particular that, if we *ignore* the second part, it is possible to obtain asymptotic bounds on the distortion in terms of the number of bits that are the same as the nonlinear approximation theory bounds in terms of the number of preserved coefficients, for classes of signals that are modeled by functions or stochastic processes in appropriate *Besov* spaces.

The reader interested in compression practice may well wonder why we keep bringing in Besov spaces (for readers unfamiliar with these spaces, we give a short primer in the Appendix). Basically, Besov spaces are a convenient mathematical setting in which one can measure smoothness of a function even if it is not uniformly smooth. For instance, a function with occasional discontinuities but smooth behavior in between (such as the realizations of the PSM) lie in a Besov space with a high smoothness index, despite the discontinuities. To a mathematician, this sounds like a good setting in which to describe images, which have edges and smoother behavior in between. However, there are many possible definitions of such general smoothness classes, and there is no *a priori* reason to select Besov spaces in particular. The answer is that Besov spaces are forced upon us *by efficient compression algorithms*: given a compression algorithm, it makes sense to derive the most general class of objects for which this algorithm exhibits a certain efficiency in a rate-distortion sense. If we think of images as functions, and if the "efficiency" of well-designed wavelet-based encoding algorithms can be measured by the asymptotic nonlinear approximation rate (an assumption that

can certainly be questioned, but which the results in this paper vindicate to some extent), then the answer is a Besov space. Certain types of Besov spaces happen to be completely characterized by the rate at which a nonlinear approximation by wavelets converges. This happy mathematical coincidence is the reason why Besov spaces (and not some other smoothness space) turn up here. In light of this, it is also not surprising that Besov spaces are indeed good at describing objects in which smoothness is not "stationary." As we shall see, they are linked to compression algorithms that effectively deal with both edges and smoother pieces.

But let us return to the total number of bits. We show that the second part of the required number of bits mentioned above, the number of addressing bits, can be *controlled only if* we add an additional ingredient. This can come from many sources. For instance, at the end of Section III we show how some extra knowledge of the approximation rate achieved by a *linear approximation* procedure using wavelets (even if it does much less well than nonlinear approximation) can be translated into an effective bound on the number of addressing bits. Interestingly, this bound turns out to dominate the number of coefficient bits, but only by a logarithmic factor. Together with a bound on the distortion in terms of the number of coefficient bits, this translates into a bound on *rate-distortion* performance that has the same functional form as the nonlinear approximation bound for this Besov space, except for an additional logarithmic factor. This fact was also recently observed in [13], for the analysis of transform coding in a (min–max) deterministic context. A close result in this direction is also given in [12]. In the case of the PSM model, we show that the logarithmic factor can be removed by exploiting more information about the process.

Although, the rate-distortion bounds found in Section III have the same asymptotic behavior as nonlinear approximation bounds, we argue that they are still unsatisfactory in terms of encoding, because the encoding strategy proposed in that section is specifically tied to a certain model, either described by a smoothness class or by the PSM. A more acceptable result would be to prove that the expected bound is achieved by an encoding strategy that is independent of the model. Two such strategies are proposed and analyzed in Sections IV and V.

In Section IV, we analyze a coding strategy that views the wavelet coefficients in each multiscale level as stemming from two different populations with different variances, where the assignment of coefficients into the two populations is done in a way that depends on the total number of bits utilized. The two populations are then coded with bit allocation rules corresponding to their two variance laws. For the PSM, as well as for a wide class of other processes, this strategy achieves a distortion bound in terms of coefficient bit rate that corresponds with the nonlinear approximation rate. Furthermore, this strategy controls the number of bits used for addressing so that it becomes negligible compared to the coefficient bit rate. This results in the desired bound on rate distortion, from which logarithmic factors have been removed.

In Section V, we analyze a tree-indexing coding strategy that is derived from a tree-structured nonlinear approximation scheme. In this case, the class of functions/stochastic processes for which the nonlinear approximation result is derived is defined slightly differently, resulting in a slightly smaller class, which still includes the PSM. Because of the tree structure in the class, the number of addressing bits is again controlled to become negligible compared with coefficient bit rate.

It is interesting to note that both of these strategies are very natural for images. The first strategy exploits the intuitive idea that the wavelet coefficients of an image can be split into two populations, which are associated with smooth and edge regions, respectively. The two populations have very different statistical properties. This point of view has recently been adopted in [3] to directly model the statistics of images. The second strategy exploits the idea that the significant wavelet coefficients of images tend to organize according to a tree structure, the main principle in the development of the algorithms in [22] and [21]. Although these encoders are expected to work well on images, a complete mathematical analysis in this context is inherently limited by the difficulties of modeling the fine properties of natural images in a rigorous way. The present paper provides a complete analysis for the simple PSM model, as well as for a versatile collection of deterministic classes. While the generalization of our analysis to more than one dimension seems difficult in the case of the PSM, it can easily be done for deterministic classes (in particular for the space BV of functions with bounded variation, which is sometimes chosen as a model for images), with natural applications to stochastic processes which live in these classes in some average sense. We devote Section VI to a discussion on the relevance and the limitations of our approach in the practical context of image compression.

Before proceeding into the main sections, a few observations involving notation and terminology are in order.

- Throughout this paper, our emphasis is on mathematical approximation results and related compression strategies based on wavelet transforms, scalar quantizers, and entropy coders. Our main efforts are geared toward deriving bounds that describe the rate-distortion performance achieved by these simple compression strategies in coding realizations of various stochastic processes. When we say rate-distortion performance of a coding strategy on a stochastic process, we mean the average distortions and the associated average rates produced by the *particular* compression strategy in coding realizations of the stochastic process (see also the discussions immediately prior to Section III-A, and at the end of Sections IV-A and IV-B for further information and generalizations). At certain places in the manuscript, we also consider the worst case distortion produced by a compression strategy using a fixed number of total bits over a class of functions.
- The bounds that we derive on rate-distortion performance enable us to explore the relationships between mathematical approximation theorems and practical compression strategies. While these bounds can also serve to upper-bound the theoretical rate-distortion functions for the stochastic processes under consideration [4], it is important to note that the rate-distortion calculations in this paper do not necessarily yield theoretical rate-distortion functions.

- This paper uses D and R to denote distortion (in $\mathcal{L}^2$ norm except where indicated) and rate (in bits), respectively. Our main results are typically of the form $D(R) \leq C\xi(R)$, where C is a constant and $\xi(R)$ is a function of R. These results are stated for functions/stochastic processes that are defined on the unit interval. As such, neither the distortion nor the rate are "normalized" (say, similar to per-pixel rate and distortion in image compression). Both quantities represent aggregate totals except where indicated otherwise.
- The calculations in this manuscript are carried out in a fashion that highlights the asymptotic rate-distortion behavior of the discussed compression strategy, i.e., we are mainly interested in how distortion is reduced as the rate in bits gets large. As such, we concentrate on the dominant terms that govern compression behavior in the asymptotic region and "absorb" the nondominant details in constants, e.g., we convert the expression $D \leq C_1 R^{-1} + C_2 2^{-R}$, for constants $C_1, C_2 > 0$, to $D \leq CR^{-1}$, where $C > 0$ is a sufficiently large constant. This approach is motivated by our objective to explore the relationships between various approximation-theoretic results (which are typically stated using the most dominant terms) and compression. The importance, relevance, and shortcomings of this analysis in real-life applications is discussed in Section VI.
- Since our results emphasize asymptotics, for notational simplicity, we ignore certain negligible overhead bits such as the number of bits required to convey an integer N (which specifies a single coding parameter), or the number of bits required to encode a single quantized scaling function coefficient. As seen in the following sections, the main portion of the required number of bits is dominated by N or even $N \log N$ bits, whereas encoding the integer N itself, for example, can be done with say $2\lceil \log N \rceil$ bits, a comparably negligible amount for large N.

II. Approximation Results

In this section, we review several results on multiscale approximation that will be exploited in the following sections to derive compression results in a rate-distortion sense. We denote by $\|\cdot\|$ the $\mathcal{L}^2$ norm, which we use systematically to measure approximation errors.

Concerning the functions that are being approximated, we distinguish between two situations.

- **The Deterministic Framework**: The functions belong to a set S that will typically be the unit ball of a certain smoothness space. For an approximation process $f \mapsto \mathcal{A}_N f$ that reduces f to N parameters, we shall study the behavior of the distortion $\varepsilon_N := \sup_{f \in S} \|f - \mathcal{A}_N f\|^2$, as N goes to $+\infty$.
- **The Stochastic Framework**: The functions are realizations of a stochastic process. We shall then study the distortion in the mean-square sense $\varepsilon_N := E(\|f - \mathcal{A}_N f\|^2)$ as N goes to $+\infty$.

In the practice of signal compression, one is often more interested in a statistical modeling and a measurement of the distortion in the mean-square sense, allowing high distortion for some rare pathological signals. Thus, our results in the following sections will mostly be formulated in this second framework. However, some of them will turn out to be obtained by a simple averaging of deterministic results.

The approximation processes that we consider rely on the multiscale decomposition with respect to a wavelet basis of compact support $\psi_{j,k}$, $j \geq 0$, $k \in K_j$. Here $K_j \subset \mathbb{Z}^d$ is a set of indexes related to the d-dimensional domain where the function is defined (one has $\#(K_j) = \mathcal{O}(2^{dj})$ for a bounded domain of $\mathbb{R}^d$). The level $j = 0$, i.e., the coarsest level, incorporates the scaling function $\varphi(x)$. Typically, $\psi_{j,k}(x)$ is defined as $\psi_{j,k}(x) = 2^{jd/2}\psi(2^j x - k)$, although this definition may have to be adapted near the boundaries of the domain under consideration. The collection of all $\psi_{j,k}$ with $j \leq J$ spans the space $\mathcal{V}_J$, and the spaces $\mathcal{V}_j$ form the multiresolution ladder $\mathcal{V}_0 \subset \mathcal{V}_1 \subset \mathcal{V}_2 \subset \cdots$. For simplicity, we consider an orthonormal basis, however, each of our results can easily be generalized to a biorthogonal wavelet basis. For a function

$$f = \sum_{j \geq 0} \sum_{k \in K_j} c_{j,k} \psi_{j,k}$$

we consider two types of approximation as follows.

- A linear approximation is defined by the projection

$$f_j = \sum_{l \leq j} \sum_{k \in K_l} c_{l,k} \psi_{l,k}$$

  i.e., keeping *the first* $N \approx 2^{dj}$ coefficients of f.
- A nonlinear approximation is defined by keeping the N *largest* coefficients of f, i.e., taking

$$\mathcal{A}_N f = \sum_{(j,k) \in E_N(f)} c_{j,k} \psi_{j,k}$$

  where the set $E_N(f)$ represents the indexes of the N largest $|c_{j,k}|$.

In Section V, we consider a more sophisticated nonlinear approximation procedure based on a *tree structure*, keeping N coefficients of f through an adaptive selection procedure that imposes additional structural constraints on the set $E_N(f)$ of retained coefficients. This strategy is related to adaptive spline approximation algorithms studied in [11], and will be particularly useful for coding purposes.

The results that we now review describe classes of deterministic sets and stochastic processes for which ε_N decays like N^{-2r}, for some given $r > 0$. Note that for more general linear and nonlinear approximation processes, e.g., polynomial, trigonometric, spline, or rational approximation, one can also define "approximation spaces" by a condition on the decay (or the summability) of ε_N (see [10, Ch. 7]). We focus here on approximation with respect to an orthonormal basis (which happens to be a wavelet basis in our case), in which case the approximation process is particularly simple since it amounts to keeping certain coefficients and discarding others.

A. Linear Approximation

The function

$$f_j = \sum_{l \leq j} \sum_{k \in K_l} c_{l,k} \psi_{l,k} \tag{2}$$

is simply the projection of f onto the space $\mathcal{V}_j$. In the deterministic framework, a classical approximation result (that can be found in [19]) is the following.

Proposition 1: Assume that the wavelet ψ has square integrable partial derivatives up to order $\mathcal{K}$, and take $r \in (0, \mathcal{K})$. Then

1) For $0 < q < \infty$, the quantities
$$\left(\sum_{j\geq 0}[2^{rj}\|f - f_j\|]^q\right)^{1/q} \tag{3}$$
and
$$\left(\sum_{j\geq 0}[2^{rj}\|f_j - f_{j-1}\|]^q\right)^{1/q} = \left(\sum_{j\geq 0} 2^{qrj}\left(\sum_{k\in K_j}|c_{j,k}|^2\right)^{q/2}\right)^{1/q}$$
are both equivalent to the norm $\|f\|_{B^r_{2,q}}$ of the function f in the Besov space $B^r_{2,q}$ ($f_{-1} = 0$).
2) For $q = \infty$, we have a similar equivalence of both
$$\sup_{j\geq 0} 2^{rj}\|f - f_j\|$$
and
$$\sup_{j\geq 0} 2^{rj}\|f_j - f_{j-1}\| = \sup_{j\geq 0} 2^{rj}\left[\sum_{k\in K_j}|c_{j,k}|^2\right]^{1/2}$$
to $\|f\|_{B^r_{2,\infty}}$.

By this equivalence we mean that there exist $C_1 > 0$ and $C_2 < \infty$, so that, for all $f \in B^r_{2,q}$
$$C_1\|f\|_{B^r_{2,q}} \leq \left(\sum_{j\geq 0}[2^{rj}\|f - f_j\|]^q\right)^{1/q} \leq C_2\|f\|_{B^r_{2,q}}$$
with similar inequalities for the other equivalences listed above.

The Besov space $B^r_{p,q}$ can be described roughly as the set of functions whose derivatives up to order r all lie in $\mathcal{L}^p$; the parameter q gives a slightly more fine-tuned description of these smoothness properties. For a primer of Besov spaces, giving the precise definitions and their basic properties, see the Appendix, which also contains more details about Proposition 1.

For $q = 2$, the Besov space $B^r_{2,2}$ reduces to the maybe more familiar Sobolev space H^r with norm defined by
$$\|f\|^2_{H^r} = \int (1 + |\xi|^{2r})|\hat{f}(\xi)|^2\, d\xi$$
where $\hat{f}$ denotes the Fourier transform of f. The equivalence stated in Proposition 1 then means that f is in H^r if and only if
$$\sum_{j\geq 0}\sum_{k\in K_j} 2^{2rj}|c_{j,k}|^2 < \infty.$$

For $q = \infty$, we obtain a slightly larger space, since $f \in B^r_{2,\infty}$ is equivalent to
$$\sum_{k\in K_j}|c_{j,k}|^2 \leq C2^{-2rj}$$
with C independent of j. Note that $f \in B^r_{2,\infty}$ is also equivalent to
$$\|f - f_j\| \leq C2^{-rj} = \mathcal{O}(N^{-r/d})$$
where we use $N \approx 2^{dj}$ terms in the linear approximation. Proposition 1 thus tells us that we can achieve a distortion on the order of $N^{-r/d}$ by linear multiscale approximation if and only if $f \in B^r_{2,\infty}$, or, roughly, if f has "r derivatives in $\mathcal{L}^2$."

In the stochastic framework, the rate of linear approximation is completely characterized by the second-order statistical properties of the signal. More precisely, assume that $f(t)$ is a stochastic process defined on $[0, 1]$, and define the autocorrelation function $\mathcal{R}(t, u) = E(f(t)f(u))$.

Definition 1: We say that $\mathcal{R}(t, u)$ is $\mathcal{C}^r$ at (t_0, u_0) if there exists a bivariate polynomial P of degree $\lfloor r \rfloor$, such that for all (t, u) in the neighbouhood of (t_0, u_0)
$$|\mathcal{R}(t, u) - P(t, u)| \leq C(|t - t_0| + |u - u_0|)^r \tag{4}$$
where $C > 0$ is a constant.[1]

The following result from [2] will be useful.

Proposition 2: Assume that the wavelet ψ has $\mathcal{M}$ vanishing moments, i.e., $\int x^k\psi(x)\, dx = 0$, $k = 0, \ldots, \mathcal{M} - 1$. Suppose that $f(t)$ is a stochastic process on $[0, 1]$, such that its autocorrelation function $\mathcal{R}(t, u)$ is uniformly $\mathcal{C}^{2r}$ on the diagonal $\{t_0 = u_0\}$, with $r < (\mathcal{M} - 1)/2$, in the sense that the constant C in (4) is independent of u, t, u_0, or t_0. Then we have
$$E(|c_{j,k}|^2) = \int \mathcal{R}(t, u)\psi_{j,k}(t)\psi_{j,k}(u)\, dt\, du \leq C2^{-(2r+1)j} \tag{5}$$
and
$$(E(\|f - f_j\|^2))^{1/2} \leq C2^{-rj}. \tag{6}$$

In order to prove (5), one simply remarks that by the vanishing moment property of ψ,
$$\int P(t, u)\psi_{j,k}(t)\psi_{j,k}(u)\, dt\, du = 0$$
where P is the polynomial in (4). The estimate (5) then follows from (4) and the support properties of $\psi_{j,k}$ using the Schwarz inequality. The bound in (6) is obtained by summing up all the variances $E(|c_{l,k}|^2)$ for bands $l > j$.

Proposition 2 easily generalizes to multivariate second-order random fields defined on bounded domains. Thus, for autocorrelation functions $\mathcal{R}$ that are sufficiently smooth near the diagonal, we obtain an expected decay in distortion via (6) within the framework of linear approximation.

In the case where the autocorrelation function is of the form $\mathcal{R}(|t - u|)$, we can give a sharper relation between the smoothness of $\mathcal{R}$ and the approximation properties of the process $f(t)$. Applying the Fourier transform, we obtain that
$$E(|c_{j,k}|^2) = \int \hat{\mathcal{R}}(\omega)2^{-j}|\hat{\psi}(2^{-j}\omega)|^2\, d\omega \tag{7}$$

[1]Typically, $P(t, u)$ consists of a truncated Taylor expansion of $\mathcal{R}(t, u)$ around (t_0, u_0).

where $\hat{\mathcal{R}}(\omega)$ and $\hat{\psi}(\omega)$ are the Fourier transforms of $\mathcal{R}$ and ψ, respectively.

Define $\hat{\Psi} := |\hat{\psi}|^2$ and let Ψ be the inverse Fourier transform of $\hat{\Psi}$. Since $\hat{\mathcal{R}}$ is a positive function, so is $\|\mathcal{R}*\Psi(2^j\cdot)\|_{\mathcal{L}^\infty}$ where $*$ denotes convolution. Viewing Ψ as a wavelet, this allows us to say that, if ψ satisfies the vanishing moments conditions in Proposition 2, and if it is smooth up to order r, then the estimate (6) is equivalent to the property that $\mathcal{R}$ belongs to the Besov space $B^{2r}_{\infty,\infty}$ (which is $\mathcal{C}^{2r}$ if $2r$ is not an integer, and a strictly larger space if $2r \in \mathbb{N}$, see the Appendix).

Note that property (6) is weaker than $E(\|f\|^2_{B^r_{2,\infty}}) < \infty$, since this would imply that $E(\sup_{j\geq 0} 2^{2rj}\|f - f_j\|^2) < \infty$. Thus, we cannot exactly relate the approximation rate (6) to a smoothness property in the mean-square sense for the general case. However, obtaining a smoothness property in a mean-square sense is possible in the following context: the quantities given by

$$E(\|f\|^2_{B^r_{2,2}}), \quad \sum_{j,k} 2^{2rj}E(|c_{j,k}|^2) \qquad \text{and} \quad E(\|f\|^2) + \sum_{j\geq 0} 2^{2rj}E(\|f - f_j\|^2)$$

are equivalent. By (7), they are also equivalent to

$$\|\mathcal{R}\|_{B^{2r}_{\infty,1}} \sim \int \hat{\mathcal{R}}(\omega)(1+|\omega|^{2r})\,d\omega.$$

Note that, in the case where $2r$ is an integer, since $\hat{\mathcal{R}}$ is positive, the finiteness of these quantities also means that $\mathcal{R}$ is $\mathcal{C}^{2r}$ in the classical sense, i.e., $2r$ times continuously differentiable.

In the case where $\mathcal{R}$ is of the form $\mathcal{R}(|t-u|)$, the approximation properties of the process in the mean-square sense are thus exactly characterized by the Besov smoothness of $\mathcal{R}$ in the scales $B^r_{\infty,q}$.

B. Nonlinear Approximation

We now consider the nonlinear approximation

$$\mathcal{A}_N f = \sum_{(j,k)\in E_N(f)} c_{j,k}\psi_{j,k} \tag{8}$$

where $E_N(f)$ is the set of indexes corresponding to the N largest coefficients, i.e., $\#(E_N(f)) = N$ and $|c_{j,k}| \geq |c_{l,m}|$ if $(j,k) \in E_N(f)$ and $(l,m) \notin E_N(f)$. Note that $\mathcal{A}_N f$ achieves the best "N-term $\mathcal{L}^2$-approximation," i.e., it minimizes $\|f - \sum_{(j,k)\in E} d_{j,k}\psi_{j,k}\|$ over all sets E of cardinality N and all possible choices of the coefficients $d_{j,k}$. In the deterministic case, a general theory was introduced in [9]. This theory relates the behavior of the best "N-term $\mathcal{L}^p$-approximation" error as N goes to $+\infty$ with the regularity of the function f. As in the linear case, Besov spaces are involved in this theory. Here, we are interested only in "N-term $\mathcal{L}^2$-approximation" and the results in [9] then have the following form.

Proposition 3: Under the same assumption on ψ as in Proposition 1, for $0 < r < \mathcal{K}$, define p by $1/p = 1/2 + r/d$. Then the quantities

$$\left[\sum_{j\geq 0}[2^{rj/d}\|f - \mathcal{A}_{2^j} f\|]^p\right]^{1/p} \tag{9}$$

and

$$\left[\sum_{N>0} N^{-1}[N^{r/d}\|f - \mathcal{A}_N f\|]^p\right]^{1/p} \tag{10}$$

are equivalent to the norm of f in the Besov space $B^r_{p,p}$.

This result and its implications are further discussed in the Appendix. It means that distortion on the order of $N^{-r/d}$ can be achieved by a nonlinear wavelet approximation, if and only if the function f has roughly "r derivatives in $\mathcal{L}^p$," where $p = \frac{2d}{d+2r}$ (and $p \neq 2$ as in the linear approximation case).

A weaker result can be obtained directly, using the equivalence (see the Appendix for details) between the norm of $B^r_{p,p}$ and the ℓ^p norm of the wavelet coefficients

$$\|f\|_{B^r_{p,p}} \approx \left[\sum_{j,k} |c_{j,k}|^p\right]^{1/p}. \tag{11}$$

Thus, if $f \in B^r_{p,p}$ and if $(\tilde{c}_n)_{n>0}$ is defined as the rearrangement of the $c_{j,k}$ in decreasing order of absolute value, then we obtain that

$$|\tilde{c}_n|^p \leq C/n \tag{12}$$

where $C > 0$ is a constant, so that

$$\|f - \mathcal{A}_N f\| = \left[\sum_{n>N} |\tilde{c}_n|^2\right]^{1/2} \leq CN^{1/2-1/p} = CN^{-r/d}. \tag{13}$$

More precisely, (13) holds if and only if the coefficients of f are in the "weak space" ℓ^p_w defined by condition (12) or by the equivalent condition[2]

$$\#\{(j,k);\, |c_{j,k}| > \varepsilon\} \leq C\varepsilon^{-p}. \tag{14}$$

The associated function space $B^{r,w}_{p,p}$ is slightly larger than $B^r_{p,p}$, and in contrast to Besov spaces, cannot be described by moduli of smoothness.

Observe that, equivalent to (12), (13), or (14) is the error estimate

$$\left\|f - \sum_{|c_{j,k}|\geq\varepsilon} c_{j,k}\psi_{j,k}\right\|^2 \leq C\varepsilon^{2-p} \tag{15}$$

which gives a characterization of nonlinear approximation rate in terms of *thresholding*. It should be noted that the spaces $B^r_{p,p}$, $1/p = 1/2 + r/d$, contain discontinuous functions for all values of r, in contrast with the spaces $B^r_{2,q}$ introduced for linear approximation, which can contain discontinuous functions only if $r \leq 1/2$ (see the Appendix). The largest r for which $f \in B^r_{p,p}$ (with $1/p = 1/2 + r/d$) tells us, in some sense, how smooth f is "in between" discontinuities. Also note that if f is in $B^r_{p,p} \cap B^\varepsilon_{2,\infty}$ for some $0 < \varepsilon < r$, then we can combine linear and nonlinear approximation results to derive the decay estimate (13) with the modified approximation

$$\mathcal{A}_N f := \sum_{(j,k)\in E_N(f),\, j<j_{\max}} c_{j,k}\psi_{j,k} \tag{16}$$

[2]The "norm" in ℓ^p_w can be defined as the infimum of $C^{1/p}$ where C is such that (12) or (14) holds. Note that this is not a true norm since it does not satisfy the triangle inequality. However, ℓ^p_w can be defined as a metric space.

where $j_{\max} = \lfloor \frac{r \log N}{\varepsilon d \log 2} \rfloor$. It is important to note that this type of restriction on the scale is necessary in many practical applications, *particularly in compression*.

If a function f is smooth except at some isolated points, one can expect that nonlinear approximation via (8) will yield a faster decay of the error than linear approximation via (2). This is because the supremum of all r such that $f \in B^r_{p,p}$, $1/p = 1/2 + r/d$, is expected to be substantially larger than the supremum of all r such that $f \in B^r_{2,\infty}$. For instance, for the univariate function f defined by $f(x) = 1$ for $\frac{1}{4} \le x \le \frac{3}{4}$, $f(x) = 0$ otherwise, one finds $f \in B^{1/2-\varepsilon}_{2,\infty}$, but $f \notin B^r_{2,\infty}$ for $r \ge 1/2$, whereas $f \in B^r_{p,p}$ for all $r < \infty$.

In the stochastic framework, one can think of such functions (in the univariate case) as processes that contain isolated jumps. In [2], a "toy model" for a piecewise-smooth 1-D process was introduced as follows.

- A finite set of locations $\{d_1, \ldots, d_L\} \subset (0, 1)$, $d_i < d_{i+1}$, is obtained as the realization on $[0, 1]$ of a Poisson process of intensity $\mu > 0$. This means in particular that the probability $P(L)$ of having L discontinuities between 0 and 1 is given by $P(L) = e^{-\mu}\mu^L/L!$, and that for a fixed number L the conditional distribution of $\{d_1, \ldots, d_L\}$ is uniform over the simplex $\{0 < x_1 < \cdots < x_L < 1\}$.
- Conditioned on fixing L and the points $d_1, \ldots, d_L$, one sets $d_0 = 0$, $d_{L+1} = 1$, and defines $f(t)$ on $[d_i, d_{i+1})$, $i = 0, \ldots, L-1$ (and $[d_i, d_{i+1}]$ for $i = L$), by $f(t) = f_i(t)$ where the f_i are independent realizations of a centered stationary process with autocorrelation function $\mathcal{R}(t, u) = \mathcal{R}(|t-u|)$, where $\mathcal{R}$ is of class $\mathcal{C}^{2r}$, $r \ge 1$. We also assume that $|f(t)| \le C$, a.e. in t, for some $C > 0$ almost surely.

In what follows, we shall refer to this model as the PSM. This type of process aims to model smooth signals disrupted by a random number of transients that can occur at arbitrary locations. The "rarity" of the transients is *hidden* in the second-order statistics. Indeed, the global autocorrelation function is given by

$$\mathcal{R}_G(t, u) = e^{-\mu|t-u|}\, \mathcal{R}(|t-u|) \tag{17}$$

and the singularity at $\{t = u\}$ simply reveals that the process is poorly regular in the mean-square sense.

From Proposition 2, it follows that the linear approximation mean-square error satisfies $\varepsilon_N \le CN^{-1}$ for some $C > 0$, since (17) indicates $\mathcal{C}^1$ (Lipschitz) type behavior for $\mathcal{R}_G(t, u)$ near the diagonal. The following results, giving different types of approximation error bounds for the PSM, were proved in [2].

Proposition 4: Any linear approximation process for the Piecewise Smooth Process (not necessarily in a wavelet basis) produces a mean-square error $\varepsilon_N \ge CN^{-1}$ for some $C > 0$. The constrained nonlinear wavelet approximation (16), with

$$j_{\max} = \left\lfloor \frac{(2r+1)\log N}{\log 2} \right\rfloor$$

yields a mean-square error $\varepsilon_N \le CN^{-2r}$ for some $C > 0$. If one replaces the wavelet basis by the Fourier basis, the nonlinear approximation error satisfies $\varepsilon_N \ge CN^{-1}$ for some $C > 0$.

This toy model is thus a typical instance of a process for which substantial gain is obtained by "switching" from linear to nonlinear approximation in the wavelet case (in contrast, no gain is obtained in the Fourier case). The above results can be obtained with more general assumptions on this model (e.g., other point processes for the discontinuities, or without stationarity). On the other hand, a "natural" generalization to the multivariate case is not clear (although recently considered in [17]).

III. From Approximation Error Bounds to Rate-Distortion Estimates

In this section, we will see how the linear and nonlinear approximation errors given in the previous section can be translated into rate-distortion based compression estimates for certain classes of stochastic processes. We will now not only retain a finite number of coefficients, but different numbers of bits will be allocated to each of these coefficients according to an appropriate strategy. Moreover, in the case of nonlinear approximation, we also have to allocate some bits to the encoding of the retained indexes. We are interested in bounding the expected error or distortion D as a function of the total number R of bits used. Note that since our models here are continuous models, i.e., the process $f(t)$ depends on the continuous (as opposed to discrete) variable $t \in [0, 1]$, R is not really a "bit rate," since we cannot normalize everything to quantities per degree of freedom or per pixel. We have infinitely many degrees of freedom in the uncompressed signal.

To derive bounds on D in terms of R from the results in the last section, we again assume that N coefficients are retained, each with a finite bit allocation only, and estimate both D and R_{coeff} (the total number of bits spent on the coefficients) as functions of N. The distortion consists of two parts: one that gives a bound on the distortion due to quantizing the N retained coefficients (typically this is ND_q, where D_q is a bound on the distortion per quantized coefficient), and another that gives the residual distortion due to the neglected coefficients (this is the same as given by the linear/nonlinear approximation results). To find the corresponding R_{coeff}, we assume that we entropy-code the retained N coefficients. The derivation of bounds on D and R_{coeff} can be done quite generally, using only *a priori* estimates on the expected Besov norm of the stochastic process. As will be seen, in order to derive bounds on $D(N)$ and $R_{\text{coeff}}(N)$ we assume that simple scalar quantization is applied, where the bit allocation for the different coefficients is bounded using estimates on the variances of the coefficients. These variance estimates follow in turn from generic Besov norm estimates. When we know more details about the process (for instance, for the PSM described at the end of Section II) we see that estimates can be sharpened and better bounds can be obtained.

In the case of linear approximation, R_{coeff} already gives the average total number of bits needed (up to an extra term to specify N itself, but this term is negligible). In the nonlinear case we also need to estimate R_{addr}, the average number of bits needed to characterize the index set of the retained coefficients, again in terms of N. The total number of bits is then $R = R_{\text{coeff}} + R_{\text{addr}}$. In both cases, the estimates on $R(N)$ and $D(N)$ can then be transformed into an estimate for $D(R)$

by the following argument. We start from $R(N) \leq R_b(N)$, $D(N) \leq D_b(N)$, where R_b and D_b are (simple) explicit functions of N. R_b is increasing and D_b is decreasing, and we assume that R_b is invertible (i.e., R_b is strictly increasing). For a *given* number of bits R, we can find the largest N such that $R_b(N) \leq R$. This is the number of coefficients that, according to our estimate, we can afford to retain within the allowed bit budget. The resulting distortion D is then bounded by

$$D \leq D_b(N) \leq D_b(R_b^{-1}(R) - 1)$$

where we have used that $R_b(N+1) > R$, and hence $N > R_b^{-1}(R) - 1$. In practice, the -1 can easily be absorbed into the constants, and we have

$$D(R) \leq C D_b(R_b^{-1}(R)).$$

Apart from the standard notation $\psi_{j,k}$ for the wavelets, it will also be convenient to use the notation g_n for the wavelets on $[0, 1]$, numbered sequentially in their "natural" order, i.e., $g_0 = \varphi$, $g_1 = \psi_{0,0}$, $g_2 = \psi_{1,0}$, $g_3 = \psi_{1,1}$, $g_4 = \psi_{2,0}$, ..., $g_{2^j+k} = \psi_{j,k}$ if $0 \leq k < 2^j$.

A. Rate-Distortion Estimates for Linear Wavelet Approximation of a Process in Some Smoothness Space

In this subsection we concentrate on the rate-distortion bounds corresponding to the *linear* approximation bounds in Proposition 1. Consider the case where the stochastic process is such that

$$E(\|f\|^2_{B^{2r}_{2,\infty}}) < \infty. \tag{18}$$

According to the definition of Besov spaces, this implies in particular that for some $C > 0$

$$E(\|f - f_j\|^2) = E\left(\left\|f - \sum_{n \leq 2^j} \langle f, g_n\rangle g_n\right\|^2\right) \leq C 2^{-2rj} \tag{19}$$

or, equivalently, that

$$\sum_{n>N} E(|\langle f, g_n\rangle|^2) \leq C N^{-2r} \tag{20}$$

for some $C > 0$. This in turn implies

$$E(|\langle f, g_n\rangle|^2) \leq C(n+1)^{-2r} \tag{21}$$

for some $C > 0$.

Let us now pursue a scalar quantization scheme, where we allocate bits following a reverse water-filling scheme [4].

For n such that $\sigma_n^2 = C(n+1)^{-2r}$ is smaller than D_q, a threshold still to be set, we allocate no bits at all, i.e., we set all coefficients beyond a certain N to zero. For $n \leq N$, we uniformly quantize $c_n = \langle f, g_n\rangle$ into $\hat{c}_n$, and entropy code each one of these "retained" coefficients.

Let $a = 2D_q^{1/2}$ be the step size of the uniform quantizer. Assuming that we use the same uniform quantizer for all the retained coefficients, we have

$$E\left(|\langle f, g_n\rangle - \hat{c}_n|^2\right) \leq (a/2)^2 = D_q. \tag{22}$$

Using (21), we have a bound on the variance of $\langle f, g_n\rangle$. In entropy-coding each $\hat{c}_n$, we use an entropy code that is optimized for a Gaussian random variable having the bounding variance σ_n^2, i.e., instead of using the optimal entropy code for each $\hat{c}_n$, which requires knowledge of the probability distribution function of $\langle f, g_n\rangle$, we use a possibly suboptimal code designed for a Gaussian random variable (having variance $\sigma_n^2 = C(n+1)^{-2r}$) undergoing uniform quantization. As shown in the derivation leading to (49) in Section IV, this uses an average bit rate

$$\ell_n \approx C' + 1/2 \log \frac{\sigma_n^2}{D_q} = C_l - 1/2 \log D_q - r \log(n+1)$$

for coefficient $\hat{c}_n$, where C', C_l are constants.

The total distortion resulting from replacing the $\langle f, g_n\rangle$, $n = 0, \ldots, N-1$ by $\hat{c}_n$, and dropping the remaining coefficients is then bounded by

$$\begin{aligned} D &= E\left(\left\|f - \sum_{n=0}^{N-1} \hat{c}_n g_n\right\|^2\right) \\ &= \sum_{k=0}^{N-1} E\left(|\langle f, g_n\rangle - \hat{c}_n|^2\right) + E\left(\sum_{n=N}^{\infty} |\langle f, g_n\rangle|^2\right) \\ &\leq N D_q + C N^{-2r} \end{aligned}$$

where $C > 0$. Choosing $D_q = C_q N^{-2r-1}$ yields[3]

$$D \leq C' N^{-2r} \tag{23}$$

for some $C' > 0$. The number of bits to convey overhead information, such as the value N, can extravagantly be absorbed into say $C_b N$, $C_b > 0$. Then, the total number of bits for overhead and for the collection $\{\hat{c}_n;\, n = 0, \ldots, N-1\}$ is

$$\begin{aligned} R &= C_b N + C_l N - \sum_{k=0}^{N-1} \left(\tfrac{1}{2} \log D_q + r \log(k+1)\right) \\ &\leq C_l' N + \sum_{k=0}^{N-1} \left(\frac{2r+1}{2} \log N - r \log(k+1)\right) \\ &= C_l' N + \frac{1}{2} N \log N - r \sum_{k=0}^{N-1} \log \frac{k+1}{N} \\ &\leq C_l' N + \tfrac{1}{2} N \log N + rN \end{aligned}$$

so that $R \leq CN \log N$ for some $C > 0$.

Together with (23) this results in

$$D(R) \leq C R^{-2r} (\log R)^{2r} \tag{24}$$

for some $C > 0$. The approximation rate in (20) is, therefore, reflected by this bound on the rate-distortion performance, except for the $(\log R)^{2r}$ factor.

Remark 5: The bound (24) is better than what would be obtained from reverse water filling arguments starting from (21), which would have lead to $D(R) \leq C R^{-(2r-1)}$ for some $C > 0$. This is because in addition to (21), we have also exploited the stronger summation estimate (20).

The following proposition summarizes our findings:

Proposition 6: Suppose that f is a stochastic process for which $E(\|f\|^2_{B^{2r}_{2,\infty}}) < \infty$. Then straightforward scalar quan-

[3] $C_q > 0$ is adjusted to yield N coefficients with $\ell_{N-1} \geq 0$.

tization and entropy coding of the wavelet coefficients of f, in their natural order, leads to the rate-distortion bound

$$D(R) \leq CR^{-2r}(\log R)^{2r} \tag{25}$$

for some $C > 0$.

Note that the same conclusion can be directly reached from (19) which is slightly weaker than $E(\|f\|^2_{B^{2r}_{2,\infty}}) < \infty$. Note also that if we had more information on the variances $E(|\langle f, g_n\rangle|^2)$, then we could improve the bound (25). For instance, if all the $E(|\langle f, g_n\rangle|^2)$ were known, then we could use this information in our entropy coder, set $\sigma_n^2 = E(|\langle f, g_n\rangle|^2)$, and allocate $l_n \approx C' - 1/2 \log D_q + \log \sigma_n$ bits to each $\hat{c}_n$, i.e., rather than using an upper bound, we could use the *exact knowledge* of the variance $E(|\langle f, g_n\rangle|^2)$. This again leads to the estimates (22) and (23), and

$$R \leq CN + \sum_{k=0}^{N-1} \left[\frac{2r+1}{2} \log N + \log \sigma_k \right]$$

for some $C > 0$. Because log is convex, we have

$$\begin{aligned} \sum_{k=0}^{N-1} \log \sigma_k &= \frac{1}{2} \sum_{k=0}^{N-1} \log \sigma_k^2 \leq \frac{1}{2} N \log \left[\frac{1}{N} \sum_{k=0}^{N-1} \sigma_k^2 \right] \\ &\leq \frac{1}{2} N \log(C_1 N^{-2r-1}) \\ &\leq -\frac{2r+1}{2} N \log N + C_2 N \end{aligned}$$

for some constants C_1 and C_2, so that

$$R \leq C_3 N$$

for some $C_3 > 0$, and

$$D(R) \leq CR^{-2r}$$

for some $C > 0$. Now the bound on R mirrors *exactly* the approximation estimate (19)! On the other hand, it may be completely unrealistic to assume that the exact σ_n^2 are known and can be used in the entropy coder. However, one can have situations where more information is available about the σ_n^2 than provided by (21).

For the PSM described at the end of Section II, we see that (17) indicates $\mathcal{C}^1$ (Lipschitz) type behavior for the global autocorrelation function $\mathcal{R}_G(t, u)$ near the diagonal. Thus, using Proposition 2 we obtain

$$E\left(|\langle f, g_n\rangle|^2\right) \leq C(n+1)^{-2}. \tag{26}$$

Mimicking the arguments above, we quantize $\langle f, g_n\rangle$ with $l_n \approx C' - \frac{1}{2} \log D_q - \log(n+1)$, resulting in

$$D \leq ND_q + CN^{-1} \tag{27}$$

for some $C > 0$, leading to the choice $D_q = C_q N^{-2}$, so that $D \leq C'' N^{-1}$ for some $C'' > 0$. Now, however, l_n for the nth coefficient is $l_n \approx C' + \log \frac{N}{n+1}$, so that

$$R \leq C'N + \sum_{n=0}^{N-1} \log \frac{N}{n+1} \leq C''' N \tag{28}$$

for some $C''' > 0$, resulting in

$$D(R) \leq CR^{-1} \tag{29}$$

for some $C > 0$. Again our better *a priori* estimate on σ_n^2 made the $\log R$ factor disappear. On the other hand, *this is not a very good rate-distortion bound*! It is dominated by the contribution from the discontinuities, which is reflected in the exponent -1 of R in the right-hand side of (29). As we will see, this bound can be improved using nonlinear approximation principles in compression.

B. Rate-Distortion Estimates for Nonlinear Wavelet Approximation of a Process in Some Smoothness Space

We now turn to the case of nonlinear approximation. Suppose that the stochastic process f is expected to lie in the weak space $B^{r,w}_{p,p}$, with $\frac{1}{p} = \frac{1}{2} + r$, i.e.,

$$E\left(\|f\|^2_{B^{r,w}_{p,p}}\right) = E\left(\|(\langle f, g_n\rangle)_{n\geq 1}\|^2_{\ell^p_w}\right) < \infty. \tag{30}$$

This implies, according the results of Section II-B, the error estimate

$$E\left(\inf_{\substack{E_N(f)\subset \mathbb{N} \\ \#E_N(f)=N}} \left\| f - \sum_{n\in E_N(f)} \langle f, g_n\rangle g_n \right\|^2 \right) \leq CN^{-2r} \tag{31}$$

for some $C > 0$.

For the PSM, we know from Proposition 4 that (31) holds with $r > 1$. According to the remarks on the space ℓ^p_w in Section II-B, the consequences of (30) can be rewritten in terms of the "rearranged decreasing coefficients" λ_k, defined by ordering the $\langle f, g_n\rangle$ in decreasing magnitude

$$\begin{aligned} n_0 &= \min\left\{ n;\ |\langle f, g_n\rangle| = \max_k |\langle f, g_k\rangle| \right\} \\ \lambda_0 &= \langle f, g_{n_0}\rangle \\ n_1 &= \min\left\{ n \neq n_0,\ |\langle f, g_n\rangle| = \max_{k\neq n_0} |\langle f, g_k\rangle| \right\} \\ \lambda_1 &= \langle f, g_{n_1}\rangle, \qquad \text{etc. } \ldots . \end{aligned}$$

Then (30) implies

$$E\left(|\lambda_k|^2\right) \leq C(k+1)^{-2r-1} = C(k+1)^{-2/p} \tag{32}$$

for some $C > 0$.

We are thus in a situation similar to before. We entropy-code the kth reordered coefficient with an average $C' + \frac{1}{p} \log N - \frac{1}{p} \log(k+1)$ bits ($D_q = C_q N^{-2/p}$), resulting in an approximate $\hat{\lambda}_k$, and

$$\begin{aligned} D &= E\left(\left\| f - \sum_{k=0}^{N-1} \hat{\lambda}_k g_{n_k} \right\|^2 \right) \\ &= E\left(\sum_{k=0}^{N-1} |\lambda_k - \hat{\lambda}_k|^2 \right) + E\left(\sum_{k=N}^{\infty} |\lambda_k|^2 \right) \\ &\leq NC_q N^{-2/p} + C'' N^{-2r} \leq CN^{-2r} \end{aligned}$$

for constants C', C_q, C'', C. On the other hand, the R_{coeff} bits we use up for $\hat{\lambda}_0, \hat{\lambda}_1, \ldots, \hat{\lambda}_{N-1}$ is bounded by

$$R_{\text{coeff}} \leq C'N + \frac{1}{p}\sum_{k=0}^{N-1} \log \frac{N}{k+1} \leq C_{rc}N \qquad (33)$$

for some $C_{rc} > 0$. This is not the total number of bits, however. We also need to encode the sequence $n_0, n_1, \ldots, n_{N-1}$. Without this information, the approximation $\sum_{k=0}^{N-1} \hat{\lambda}_k g_{n_k}$ of f *cannot be reconstructed*.

In order to estimate the number of bits R_{addr} needed to encode the $n_0, \ldots, n_{N-1}$, we need some information on the possible distribution of the large values of $|\langle f, g_n\rangle|$. This information is not contained in (30) or (31), which only tells us how fast the rearranged coefficients decay, not what their index was prior to rearrangement. So we must obtain it elsewhere.

If we had bounds on each n_k, then we could write an upper bound for the number of addressing bits. A bound of this nature is provided by *linear* approximation. For instance, suppose that, in addition to the nonlinear approximation rate (31), we also have a bound for the linear approximation rate

$$E\left(\left\|f - \sum_{n=0}^{N-1} \langle f, g_n\rangle g_n\right\|^2\right) \leq CN^{-2\gamma} \qquad (C>0) \quad (34)$$

for some $\gamma > 0$. Then, defining $M(N) = \lceil N^{r/\gamma}\rceil$, we obtain that

$$E\left(\left\|f - \sum_{n=0}^{M(N)} \langle f, g_n\rangle g_n\right\|^2\right) \leq CN^{-2r}. \qquad (35)$$

Therefore, if we modify the above bit allocation in the sense that we allocate no bits to λ_k if $n_k > M(N)$, we still find that the distortion satisfies

$$D = E\left(\left\|f - \sum_{k\leq N-1,\, n_k \leq M(N)} \hat{\lambda}_k g_{n_k}\right\|^2\right) \leq CN^{-2r} \quad (36)$$

for some $C > 0$.

Since there are at most N indexes to encode among $M(N) = \lceil N^{r/\gamma}\rceil$ possible values, we derive that the number of addressing bits is controlled by

$$R_{\text{addr}} \leq C_{ra} N \log N. \qquad (37)$$

Comparing with (33), we see that R_{addr} *dominates* R_{coeff}! Combining (37) with (33) and (36), this leads again to $\log R$ factors in the rate-distortion estimate as summarized by the following proposition.

Proposition 7: Suppose that f is a stochastic process for which $E(\|f\|^2_{B^{r,\,w}_{p,\,p}}) < \infty$, where $\frac{1}{p} = \frac{1}{2} + r$. Then, the corresponding *a priori* bounds on the variances of the reordered wavelet coefficients lead to the bounds

$$D(N) \leq CN^{-2r} \quad \text{and} \quad R_{\text{coeff}}(N) \leq CN$$

if we retain N coefficients. If, in addition, $E(\|f\|^2_{B^{\gamma}_{2,\,\infty}}) < \infty$ for some $\gamma > 0$, then we can control the number of addressing bits by

$$R_{\text{addr}}(N) \leq CN \log N$$

resulting in

$$D(R) \leq CR^{-2r}(\log R)^{2r} \qquad (38)$$

for some $C > 0$.

Note that in the whole derivation, we have essentially used (31) and (32) which are slightly weaker than (30). Can we do better if we have more information on the stochastic process? Ideally, we would like to remove the $\log R$ factors from the estimate (38). Let us consider the PSM again. Since both (31) and (34) hold for this model (with $\gamma = 1/2$), we certainly have (38). If we can improve (37) (so that the upper bound on $R(N)$ becomes linear in N) then the $\log R$ factors will disappear from the rate-distortion bound. To achieve this, we propose the following strategy, a slight variant on what was used in [2] to prove Proposition 4.

Suppose that we retain N coefficients. For a given realization f, we have L discontinuities, located at $d_1, \ldots, d_L$.

Case 1: Suppose that

$$L \leq L_0 = \left\lfloor \frac{N}{4rW \log N} \right\rfloor$$

where $W = \text{supp}\,(\psi)$ is the width of the compact mother wavelet support (if $L > L_0$, then we will follow a different plan which is examined later). At every level j we find the indexes $k \in K_j$ for which $\text{supp}\,(\psi_{j,k}) \cap \{d_1, \ldots, d_L\} = \emptyset$. We call these the indexes "of the first kind," and we denote this by $k \in K_j^1$. The other $k \in K_j$ are then "of the second kind," $k \in K_j^2 = K_j \backslash K_j^1$.

We split N into N_1 and N_2, where $N_1 + N_2 = N$, and choose $N_2 = \lfloor 2rWL \log N\rfloor$. Since $L \leq L_0$, we have $N_1 = N - N_2 \geq N/2$. In each group (first kind or second kind), we propose to retain the first N_ℓ ($\ell = 1, 2$) coefficients $\langle f, \psi_{j,k}\rangle$, $k \in K_j^\ell$, ordered in their natural order. This fills J_1 levels of coefficients of the first kind, and J_2 levels of coefficients of the second kind. We have

$$\log N \geq J_1 \geq \lfloor \log N_1\rfloor \geq \lfloor \log N\rfloor - 1$$
$$N_2 \geq J_2 > \left\lfloor \frac{N_2}{WL} \right\rfloor \geq \lfloor 2r \log N\rfloor.$$

where the second inequality follows from the compact support properties of the wavelet basis.[4] In order to characterize the sets K_j^ℓ ($\ell = 1, 2$) precisely for all $j \leq J_1 < J_2$, we thus need to encode L, as well as the locations $d_1, \ldots, d_L$ up to precision 2^{-J_2}.

Using $\log L \leq C_1 \log N$ for some $C_1 > 0$, the overhead bits to encode the value L as well as other coding parameters (N,

[4]At level j there are less than $supp(\psi_{j,k})/2^{-j} = 2^{-j}\text{supp}\,(\psi)/2^{-j}$ wavelet coefficients overlapping a single-point discontinuity. Hence, there are no more than $W = \text{supp}\,(\psi)$ wavelet coefficients overlapping a single-point discontinuity in *any* level. With L discontinuities there at most WL wavelet coefficients overlapping point discontinuities in any level.

etc.) can be absorbed into $C_2 \log N$ for some $C_2 > 0$. We thus require no more than

$$C_2 \log N + J_2 L \le C_a L^2 \log N \tag{39}$$

bits for some $C_a > 0$, in order to encode L and the locations $d_1, \ldots, d_L$ up to precision 2^{-J_2}. Once we have spent this number of bits, we know exactly which are the coefficients of the first/second kind.

Let $\tilde{E}$ denote conditional expectation given that L is pinned down, as well as the $d_1, \ldots, d_L$ with precision 2^{-J_2}. Now for $k \in K_j^1$

$$\tilde{E}\left(|\langle f, \psi_{j,k}\rangle|^2\right) \le C 2^{-j(2r+1)} \tag{40}$$

which follows from Proposition 2 since the conditional autocorrelation function for two points not separated by discontinuities is $\mathcal{C}^{2r}$. For $k \in K_j^2$

$$\tilde{E}\left(|\langle f, \psi_{j,k}\rangle|^2\right) \le C 2^{-j} \tag{41}$$

for some $C > 0$, since in this case we always have

$$\tilde{E}(|\langle f, \psi_{j,k}\rangle|^2) \le E\left(\int_{\mathrm{supp}\,(\psi_{j,k})} f(t)^2\, dt\right) \le C 2^{-j}$$

by Schwarz inequality. We can use this to allocate bits: $C' - 1/2 \log D_q - (r+1/2)j$ for the coefficients of the first kind, $C'' - 1/2 \log D_q - j/2$ for those of the second kind, where C' and C'' are constants. This leads to an expected distortion bounded by

$$\begin{aligned}\tilde{D}(N) &\le N D_q + C 2^{-2rJ_1} + C L 2^{-J_2} \\ &\le N D_q + C''' L N^{-2r}\end{aligned} \tag{42}$$

where C, C''' are constants and the $\tilde{}$ again indicates that we are working with conditional expectations. As before, this suggests the choice $D_q = C_q N^{-2r-1}$, leading to

$$\tilde{D} \le C L N^{-2r}, \qquad \text{for some } C > 0. \tag{43}$$

The corresponding bit cost R_{coeff} is then bounded by

$$\begin{aligned}\tilde{R}_{\mathrm{coeff}}(N) &\le C_c N + \sum_{j=0}^{J_1} \sum_{k \in K_j^1} [(r+1/2)\log N - (r+1/2)j] \\ &\quad + \sum_{j=0}^{J_2} \sum_{k \in K_j^2} [(r+1/2)\log N - j/2] \\ &\le C_c N + \sum_{j=0}^{J_1} 2^j (r+1/2)(J_1 + 2 - j) \\ &\quad + \sum_{j=0}^{J_2} LW[(r+1/2)\log N - j/2] \\ &\le C_c N + C_s 2^{J_1} + C_l L J_2 \log N \\ &\le C_{rc} N + C'_{rc} (L \log N)^2\end{aligned}$$

for constants C_c, C_s, C_l, C_{rc}, and C'_{rc}. Note that all of this is for $L \le L_0$.

Case 2: If $L > L_0$, then we simply give up allocating bits: we put all the coefficients to zero, not spending any bits. The resulting expected distortion is then bounded by

$$\tilde{\tilde{D}}(N) \le E(\|f\|^2) \le \mathcal{R}(0). \tag{44}$$

To obtain the full expected distortion and bit cost, we must now sum our conditional estimates, weighted according to their probability. Integrating over the possible locations of the $d_1, \ldots d_L$, for fixed L, just contributes a constant factor, so we need only check the summation over L. We find

$$\begin{aligned}D(N) &\le C N^{-2r} \sum_{L=0}^{\lfloor N/(4rW\log N)\rfloor} L \frac{e^{-\mu}}{L!}\, \mu^L \\ &\quad + \mathcal{R}(0) \sum_{L > \lfloor N/(4rW\log N)\rfloor} \frac{e^{-\mu}}{L!}\, \mu^L \\ &\le C N^{-2r} \mu + \mathcal{R}(0) \frac{\mu^{(\lfloor N/(4rW\log N)\rfloor + 1)}}{(\lfloor N/(4rW\log N)\rfloor + 1)!} \\ &\le C_d N^{-2r}\end{aligned}$$

where we have used $\mu = \sum_L L \frac{e^{-\mu}}{L!}\, \mu^L < \infty$, and

$$\begin{aligned}R(N) &= R_{\mathrm{coeff}}(N) + R_{\mathrm{addr}}(N) \\ &\le C_{rc} N + C'_{rc} (\log N)^2 \sum_{L=0}^{\lfloor N/(4rW\log N)\rfloor} L^2 \frac{e^{-\mu}}{L!}\, \mu^L \\ &\quad + C_a (\log N) \sum_{L=0}^{\lfloor N/(4rW\log N)\rfloor} L^2 \frac{e^{-\mu}}{L!}\, \mu^L \\ &\le C_r N\end{aligned}$$

where we have used $\sum_L L^2 \frac{e^{-\mu}}{L!}\, \mu^L < \infty$. This then leads to the desired bound

$$D(R) \le C R^{-2r}, \qquad \text{for some } C > 0. \tag{45}$$

Recall that the "competing" bound based on linear approximation is $D(R) \le C R^{-1}$ (see (29)). It is important to note that while the competing bound is determined by the global autocorrelation $\mathcal{R}_G$ and its poor overall smoothness, the bound we have just derived takes advantage of local smoothness $(r > 1)$. Note also that this is the same asymptotic behavior as that would be obtained had we considered a stochastic process with an autocorrelation function in $\mathcal{C}^{2r}$, *without any discontinuities thrown in*. In that case, for all j, k one has

$$E\left(|\langle f, \psi_{j,k}\rangle|^2\right) \le C 2^{-j(2r+1)} \tag{46}$$

or

$$E\left(|\langle f, g_n\rangle|^2\right) \le C n^{-(2r+1)} \tag{47}$$

and the bound (38) follows from arguments similar to the derivation of (29). *So our strategy "erases" the effect of the discontinuities in the rate-distortion bound, just like in the nonlinear approximation results in [2].*

However, although (45) is the desired result, we really obtained it by cheating: our strategy subsumed an *exact* knowledge of the nature of the stochastic process in order to create the two classes K_j^1 and K_j^2, and it used the fact that coefficients of the second kind all "line up" underneath the discontinuities

in order to get away with a *very low number of addressing bits* (with $R_{\text{addr}} \leq C \log N$). This is quite unfair: it means that we inspect every realization in order to decide on the number and location of the singularities, and classify the N wavelet coefficients with perfect accuracy. It would be much more interesting to see whether we can improve on (37) by using reasonable, "real-life" coding strategies that use only the wavelet coefficients of the realization, and that do not assume a detailed knowledge of the stochastic process itself. This is the topic of the next two sections.

IV. Coding a Mixture Distribution of Two Multiresolution Populations

In this section, we construct and analyze a real-life coding algorithm that is motivated by the PSM process but, unlike the coder sketched at the end of the previous section, is not structured around the specific parametric form of the PSM. Instead, this coder models the process as a mixture distribution of two populations of wavelet coefficients: those describing smooth regions of the process, and those contributing to the description of singularities in the process. At the end of the previous section, we established bounds on the variance of these two populations with the coefficients of the first kind satisfying

$$\tilde{E}\left(|\langle f, \psi_{j,k}\rangle|^2\right) \leq C 2^{-j(2r+1)}$$

and those of the second kind satisfying

$$\tilde{E}\left(|\langle f, \psi_{j,k}\rangle|^2\right) \leq C 2^{-j}$$

where the expectations are conditioned as before but without extra constraints like $L \leq L_0$. The coder outlined in this section uses a practical addressing scheme for associating the wavelet coefficients in each band with one of these two populations, and applies entropy-coded, uniform-stepsize scalar quantizers for coding the coefficients in each band. Our objective is to upper-bound the performance of this practical coder in coding the PSM process. We develop this coder in two stages. In the first stage, we develop a coding syntax designed to code the PSM process efficiently. To simplify analysis of the performance achievable by the syntax, we allow the encoder to use ideal knowledge of the input process (e.g., number of edges, location of edges, etc.) in its encoding strategy. In the second stage, we show that the performance of a practical encoding strategy that does not use ideal knowledge also satisfies the same bound.

Since all random variables are characterized by bounds on their variances, we first upper-bound the number of bits and distortion needed for coding an arbitrary random variable whose variance is bounded by σ_X^2 using a uniform-step-size scalar quantizer followed by entropy coding. The entropy coder is optimized for a Gaussian random variable with variance equal to σ_X^2. A uniform-step-size scalar quantizer with step size a bounds the quantization error $-a/2 < x - Q(x) \leq a/2$. To simplify later analysis in this section, we define $\hat{D}_q = \frac{a^2}{4}$, which serves to bound the quantizer distortion

$$D_q \leq \hat{D}_q.$$

It will be convenient to use $\hat{D}_q$ in much of the analysis since it is deterministically related to the quantizer step size a, while upper-bounding the real quantizer distortion for any input random variable.

For an arbitrary zero-mean random variable Y, with density $f_Y(y)$ and variance $\sigma_Y^2 < \infty$, and a zero-mean Gaussian random vriable X with the variance $\sigma_X^2 \geq \sigma_Y^2$, the number of bits needed for coding the quantized variable $\hat{Y}$ using the entropy coder designed for X is given by[5]

$$\begin{aligned} H(\hat{Y}) &= \sum_i P_Y(i) \log(1/P_X(i)) \\ P_Y(i) &= \int_{(i-1/2)a}^{(i+1/2)a} f_Y(x)\, dx \\ P_X(i) &= \int_{(i-1/2)a}^{(i+1/2)a} f_X(x)\, dx \\ &= f_X(x_i) a \\ &= \frac{1}{\sqrt{2\pi\sigma_X^2}}\, e^{-x_i^2/(2\sigma_X^2)} a \end{aligned}$$

where the next to last step is by the mean value theorem, and $(i-1/2)a \leq x_i \leq (i+1/2)a$. We have

$$\begin{aligned} H(\hat{Y}) &= \sum_i P_Y(i) \log\left(\sqrt{2\pi\sigma_X^2}\, e^{x_i^2/(2\sigma_X^2)}\, a^{-1}\right) \\ &= 1/2 \log(2\pi\sigma_X^2) - \log(a) + \log(e) \sum_i P_Y(i) x_i^2/(2\sigma_X^2). \end{aligned}$$

Since $\sigma_X^2 \geq \sigma_Y^2$

$$\begin{aligned} \sigma_X^2 &\geq \sum_i P_Y(i) E\left[Y^2 | (i-1/2)a \leq Y \leq (i+1/2)a\right] \\ &= \sum_i P_Y(i) y_i^2 \end{aligned}$$

where $(i-1/2)a \leq y_i \leq (i+1/2)a$. Then, since y_i and x_i are constrained to the same intervals, for all $i \neq 0$, $(3y_i)^2 \geq x_i^2$, and

$$\begin{aligned} H(\hat{Y}) &\leq 1/2 \log(2\pi\sigma_X^2) - \log(a) \\ &\quad + \log(e) \frac{\sum_{i \neq 0} P_Y(i) x_i^2}{2 \sum_i P_Y(i) y_i^2} + \log(e) P_Y(0) x_0^2/(2\sigma_X^2) \\ &\leq 1/2 \log(2\pi\sigma_X^2) - \log(a) \\ &\quad + 9 \log(e)/2 + \log(e) x_0^2/(2\sigma_X^2). \end{aligned}$$

For $a \leq 2\sigma_X$, the last term on the right is bounded by $\log(e)/2$, and the desired bound on the number of bits is determined as

$$H(\hat{Y}) \leq C + \log(\sigma_X) - \log(a)$$

where C is a constant. For $a > 2\sigma_X$, we let $a = K\sigma_X$, with $K > 2$. By the definition of x_0

$$\begin{aligned} 1 \Big/ \left(\sqrt{2\pi}\sigma_X\right)\, e^{-x_0^2/(2\sigma_X^2)} K\sigma_X &\geq \int_{-\sigma_X}^{\sigma_X} f_X(x)\, dx = C_1 \\ \log(e) x_0^2/(2\sigma_X^2) &\leq \log(K) + C_2 \end{aligned}$$

[5]When encoding discrete random variables, it is well known that the average number of bits output by practical entropy coders (say, by utilizing arithmetic codes) can be made arbitrarily close to the theoretical minimum given by the entropy. We will hence use the entropy expression directly to compute the required number of bits.

where C_1, C_2 are constants, giving

$$H(\hat{Y}) \leq 1/2\log(2\pi\sigma_X^2) - \log(K\sigma_X) + \log(K) + C_3 = C_4$$

for constants C_3, C_4. Thus, for all a, $H(\hat{Y})$ is bounded with constants C and C'

$$H(\hat{Y}) \leq \max(C,\, C' - \log(a) + \log(\sigma_X)). \tag{48}$$

We note that in applying reverse water-filling for developing the algorithm that follows, all scalar quantizers applied to random variables with variances less than σ_X use step sizes $a \leq 2\sigma_X$, letting us write

$$\begin{aligned} H(\hat{Y}) &\leq C' - \log(a) + \log(\sigma_X), \\ &\leq C + 1/2\log\left(\frac{\sigma_X^2}{\hat{D}_q}\right). \end{aligned} \tag{49}$$

A. Rate-Distortion Bounds for Quantizing Piecewise-Smooth and Edge Wavelet Coefficients

We now develop and analyze the rate-distortion performance of a wavelet-based coding syntax suitable for coding the PSM process. This section focuses on defining a practical syntax for quantizing and transmitting wavelet coefficient values, but does not require that the encoder use realistic strategies for encoding within that syntax. In fact, the encoding strategies analyzed in this section assume that the encoder has perfect knowledge of features of each realization of the process: e.g., the number of edges, exact location of edges. The next section replaces these strategies with a realistic encoder.

We begin by analyzing the rate-distortion characteristics of a uniform-step-size scalar quantizer applied to a random process having a "smooth," stationary autocorrelation function $\mathcal{R}(\tau)$ which is $\mathcal{C}^{2r}$ at $\tau = 0$.

Suppose we keep the first N wavelet coefficients over such a process, where for convenience, we assume $N = 2^{J_s}$. We know from Proposition 4 that the average distortion we incur is no more than

$$D_s \leq CN^{-2r} + N\hat{D}_q$$

where $\hat{D}_q$ bounds the quantization distortion for each coefficient that we keep, assuming that all scalar quantizers use the same step size.

In the jth band ($j \leq J_s$), there are 2^j coefficients and

$$E[|c_{j,k}|^2] \leq C_1 2^{-(2r+1)j} \tag{50}$$

from Proposition 2. Using (49), in the jth band, the average total number of bits is no more than

$$\begin{aligned} &\leq \frac{1}{2}\, 2^j \log_2\left(\frac{C_b 2^{-(2r+1)j}}{D_q}\right) \\ &= 2^{j-1}\log_2\left(\frac{C_b}{\hat{D}_q}\right) - (2r+1)j2^{j-1} \end{aligned}$$

where C_b is a constant. Thus, summing over all bands ($j \leq J_s$),[6] we get

$$\begin{aligned} R_s \leq \frac{1}{2}\log_2\left(\frac{C_b}{\hat{D}_q}\right)(2^{J_s+1} - 1) \\ -(2r+1)((J_s+1)2^{J_s} - 2^{J_s+1} + 1). \end{aligned} \tag{51}$$

Now, because the variance bounds decrease over bands and we retain all the bands $j \leq J_s$, for water-filling on the bound we set the quantizer step sizes so that

$$\hat{D}_q = C_b 2^{-(2r+1)(J_s)}. \tag{52}$$

Equation (51) becomes

$$\begin{aligned} R_s &\leq (2r+1)(J_s)\left(2^{J_s} - \frac{1}{2}\right) \\ &\quad - (2r+1)[(J_s+1)2^{J_s} - 2^{J_s+1} + 1] \\ &= (2r+1)2^{J_s} - \frac{(2r+1)J_s}{2} - (2r+1). \end{aligned}$$

Using $N = 2^{J_s}$, it follows that $R_s \leq C_s N$. Using (52), the distortion bound becomes

$$\begin{aligned} D_s &\leq CN^{-2r} + N\hat{D}_q \\ &= C'N^{-2r} \\ D_s(R_s) &\leq C_s' R_s^{-2r}. \end{aligned} \tag{53}$$

Now consider the rate-distortion performance of wavelets coding edges over a *uniform* background. For example, this would be the case if the stationary autocorrelation function $\mathcal{R}(\tau)$ determining the correlation of the smooth patches in the PSM were a constant ($\mathcal{R}(\tau) = \mathcal{R}(0)$), resulting in a piecewise-constant process. For the moment, we pretend that we know the locations of these edges. Consider first the case when there is one edge over a uniform background. At band j, the total number of coefficients overlapping a given edge is at most C_o, i.e., an edge over a uniform background affects at most a *constant* number of wavelet coefficients in each band as the utilized wavelet basis has compact support. From the bound in (41) we know that the wavelet coefficient variances over the edge decay no worse than

$$E\left[|c_{j,k}|^2\right] \leq C_b 2^{-j} \tag{54}$$

where j indexes the coefficients over bands. Suppose we keep J_e bands. Over the edge we have a total distortion

$$\begin{aligned} D_e &\leq C_o J_e \hat{D}_q + C_o C_b \sum_{i=J_e+1}^{\infty} 2^{-i} \\ &\leq C_o J_e \hat{D}_q + C_o C_b 2^{-J_e} \end{aligned}$$

where $\hat{D}_q$ bounds the quantization distortion on the transmitted coefficients (assuming equal step sizes) and $C_o C_b 2^{-J_e}$ is distortion due to dropped coefficients.

Now using bound (54) we have

$$\begin{aligned} R_e &\leq C_o \sum_{i=0}^{J_e} \frac{1}{2}\, \log_2\left(\frac{C_b 2^{-i}}{\hat{D}_q}\right) \\ &= C_o \frac{1}{2}(J_e+1)\left(\log_2\left(\frac{C_b}{\hat{D}_q}\right) - \frac{J_e}{2}\right) \end{aligned}$$

which is the number of bits required to transmit the coefficients with quantization distortion bounded by $\hat{D}_q$. Because we retain

[6]For notational convenience, we neglect the number of bits required for the scaling function coefficient.

only J_e bands, for reverse water-filling we set stepsizes such that

$$\hat{D}_q = C_b 2^{-J_e} \tag{55}$$

and we have

$$R_e \leq C_o \frac{1}{2}(J_e+1)\left(\frac{J_e}{2}\right) \tag{56}$$

$$D_e \leq C_o (J_e+1) 2^{-J_e} C_b. \tag{57}$$

From (56), $J_e \geq 2\sqrt{\frac{R}{C_o}} - 1$, and because (57) is decreasing with respect to J_e, we have that the rate-distortion bound for coding a single *fixed* edge with wavelets is

$$D_e(R_e) \leq C' \sqrt{R_e}\, 2^{-2\sqrt{\frac{R_e}{C_o}}} \tag{58}$$

for some constant C'.

Now assume we have $\kappa > 0$ edges for which we know the locations. Having κ edges instead of one edge modifies C_o, the number of coefficients over edges, to $C_o\kappa$. Some coefficients may lie over more than one edge but note that the bound (54) is loose enough to accommodate this case. Equations (56) and (57) are modified to

$$R_e \leq C_o \kappa \frac{1}{2}(J_e+1)\left(\frac{J_e}{2}\right) \tag{59}$$

$$D_e \leq C_o \kappa (J_e+1) 2^{-J_e} C_b \tag{60}$$

and (58) becomes

$$D_e(R_e) \leq C'' \sqrt{R_e \kappa}\, 2^{-2\sqrt{\frac{R_e}{C_o \kappa}}} \tag{61}$$

where C'' is a constant independent of the number of edges κ.

Let us now calculate the number of bits required to convey the "addresses" of wavelet coefficients over edges. Again, assume there are κ edges.

At the jth band there are 2^j locations and at most $C_o\kappa$ of these are over edges. Clearly, we can specify the addresses of the coefficients over edges with j bits per coefficient. In our addressing scheme, we transmit the location of each edge coefficient in the jth band with $j+1$ bits, with the understanding that a say $(j+1)$th bit *zero* is reserved for an ⟨*end of address information in the jth band*⟩ symbol. This way, the decoder will know that the transmission of address information for the jth band is completed and can start decoding the address information for the next band, and so on. Assuming that a total of J_e bands is kept, it can be seen that the total bit cost of specifying the locations of coefficients over edges in such a fashion results in no more than R_L bits, where

$$\begin{aligned} R_L &= \sum_{j=0}^{J_e} (C_o\kappa + 1)(j+1) \\ &= (C_o\kappa+1)(J_e/2+1)(J_e+1) \end{aligned} \tag{62}$$

where $C_o\kappa + 1$ reflects the ⟨*end of address information in the jth band*⟩ symbol.

Now the above results will be combined to bound the rate-distortion performance of wavelets over a piecewise-constant edgy process. Adding the location bits R_L from (62) to (59) we now have a modified number of bits and

$$R_e \leq (C'\kappa)\left((J_e+1)\frac{J_e}{2}\right)$$

$$D_e \leq C''\kappa (J_e+1) 2^{-J_e}.$$

This only changes constants and we have

$$D_e(R_e, \kappa) \leq C_1 \sqrt{R_e \kappa}\, 2^{-C_2\sqrt{\frac{R_e}{\kappa}}} \tag{63}$$

where $C_1 > 0$ and $C_2 > 0$ are constants and notation is changed to reflect the dependence of our rate-distortion curve on the number of edges.

To obtain the overall $D(R)$ for both smooth and edge coefficients, we force equal-step-size quantizers, thus relating R_e, R_s, D_s, and D_e through the expressions for $\hat{D}_q$. Specifically, we have

$$\hat{D}_q = C' 2^{-J_s(2r+1)} = C''' 2^{-J_e} \tag{64}$$

from which $J_e = J_s(2r+1) + C$. Observe that asymptotically $J_e \geq J_s$ irrespective of C. Using this relation, the average number of bits and distortion for the smooth process are

$$R_s \leq C_1 2^{J_s} \quad \text{and} \quad D_s \leq C_1' 2^{-J_s(2r)}$$

and those for the edge process, conditioned on having κ edges, are

$$\begin{aligned} R_e &\leq C_2 \kappa (J_s(2r+1)+1) J_s(2r+1) \quad \text{and} \\ D_e &\leq C_2' \kappa (J_s(2r+1)+1) 2^{-J_s(2r+1)} \\ &\leq C_2'' \kappa 2^{-J_s(2r)}. \end{aligned}$$

We recall that our encoding strategy uses an oracle who provides us the correct value of κ for every realization, as well as the exact location of each edge, which we use to perfectly identify all coefficients of the edge process.

To get bounds on the overall average number of bits and distortion we average over the random variable κ

$$\begin{aligned} R &= R_s + E_\kappa[R_e] \leq C' 2^{J_s} + C'' E[\kappa](J_s(2r+1)+1)^2 \\ D &= D_s + E_\kappa[D_e] \leq K' 2^{-J_s(2r)} + K'' E[\kappa] 2^{-J_s(2r)} \end{aligned}$$

where C', C'', K', and K'' are constants. Assuming that $E[\kappa]$ is a finite parameter of the PSM process, and can be incorporated into a constant, we get

$$\begin{aligned} R &\leq C''' 2^{J_s} \\ D &\leq K''' 2^{-J_s(2r)} \end{aligned} \tag{65}$$

for constants C''' and K''', resulting in the desired

$$D \leq D_B(R) = C_b R^{-2r}. \tag{66}$$

We note that our analysis accounts for the actual number of edge coefficients coded, but overcounts the number of smooth coefficients that are coded. By doing so, we overestimate the number of bits needed to code the smooth coefficients, and overestimate the distortion incurred in coding them. This is consistent with an upper bound on the rate-distortion curve.

Finally, we note that the number of bits used by the coding strategy described above is a random variable whose expected value is analyzed in (65). Two factors contribute to this random number of bits: the use of variable-length coding on the wavelet coefficients, and the variable number of edges found in each realization of the process. Virtually all source coders used in practice today employ some form of variable-length coding, motivating us to consider entropy coding in this coder. Together with the variable number of edges, we can expect that the number of bits will depend on the complexity of the input realization.

However, if a deterministic constraint on the number of bits is specified, the step size a (along with J_s and J_e through (52) and (55)) can be used to adjust both R_s and R_e to meet the specified constraint for each realization. For the coder analyzed in this section, any adjustment on the number of bits involves knowing the value of κ for each realization, but the modified coding strategy described in the following section permits adjustment without explicitly knowing κ. It is worth noting that the number of bits produced by standard image coding algorithms like JPEG [16] also varies with the complexity of the input image, and the number of bits used for coding any particular image is adjusted through a "quality factor" parameter that adjusts quantization step sizes.

B. A Practical Encoding Strategy

The coding syntax developed in the previous subsection is composed of the following components.

- Overhead bits defining constants (e.g., r) and quantizer step size a. From these variables, J_s and J_e can be determined, as well as the bounding variances σ_X for each of the populations of each band $j \leq J_e$. The bounding variances fully specify all entropy codes used by the algorithm.
- A sequence of addresses, for each band $j \leq J_e$, terminated by a fixed code. For $j \leq J_s$, the sequence separates coefficients into two classes. Both classes are quantized with the same uniform step-size scalar quantizer, but the two classes use different probability models for entropy-coding the quantizer outputs. For $J_s < j \leq J_e$, those coefficients addressed by the sequence are coded, while all others are not coded.
- A bit stream representing the entropy-coded outputs of the scalar quantizers.

The coder of the previous section is informed by an oracle about the number and precise location of edges in the PSM, and uses this information for encoding the PSM within the above syntax. This section shows that, without the help of an oracle, a simple practical encoding strategy can be applied to the same coding syntax to satisfy the same performance bound (66).

Given a step size a, the bounds from (65) can be expressed in terms of a via (64) to individually bound the number of bits and distortion

$$\begin{aligned} R &\leq R_a = C_r a^{-\frac{1}{(r+1/2)}} \\ D &\leq D_a = C_d a^{\frac{2r}{(r+1/2)}}. \end{aligned} \tag{67}$$

These individual bounds satisfy a linear bound on the rate-distortion performance that is stronger than the $D_B(R)$ bound in (66). Since $D_B(R)$ is a convex function of R, the supporting line $D_L(R)$ at any R_a

$$D_L(R) = \lambda_a(R - R_a) + D_B(R_a), \quad \text{for } \lambda_a = \frac{d}{dR} D_B(R)|_{R=R_a}$$

satisfies $D_L(R) \leq D_B(R)$, while bounding from above any R, D pair satisfying (67). The slope λ_a of this line is given by

$$\begin{aligned} \lambda_a &= C_b(-2r)R_a^{-(2r+1)} \\ &= -Ca^2. \end{aligned}$$

Note that λ_a is always negative. Our practical encoding strategy allows both the number of bits and distortion to change from those of the oracle-based encoder, but guarantees that, for a given step size a, the rate-distortion pair (R_p, D_p) achieved by the practical coder is at least as far below the bounding line $D_L(R)$ as the pair (R_o, D_o) achieved by the oracle-based coder, i.e.,

$$D_p - \lambda_a R_p \leq D_o - \lambda_a R_o. \tag{68}$$

A sufficient condition for the average distortions and numbers of bits produced by the two encoding strategies to satisfy (68) is that a similar inequality is satisfied pointwise in the encoding of each and every wavelet coefficient. Namely, if $e_p(j, k)$, $e_o(j, k)$, $r_p(j, k)$, and $r_o(j, k)$ are the errors incurred and number of bits used in coding the j, k wavelet coefficient by each of the two encoding strategies, then, for every j, k

$$e_p(j, k) - \lambda_a r_p(j, k) \leq e_o(j, k) - \lambda_a r_o(j, k). \tag{69}$$

The practical encoding strategy is designed to satisfy this condition.

In designing the practical coding strategy, we assume that the smoothness parameter r and constants of the process are either known explicitly or estimated from the process during a training period. From these parameters, the bounding variances of the smooth and edge coefficient populations of each band are determined. These variances, along with a quantizer step size, determine both the values of J_s and J_e, and the entropy codes to be used for all quantized variables. To replace the oracle, our practical algorithm must assign each coefficient in bands $j \leq J_s$ to either the smooth or edge populations, without knowing either the number or location of edges. We define different assignment rules for the two cases: 1) $j \leq J_s$ and 2) $J_s < j \leq J_e$.

Case 1: $j \leq J_s$: For these bands, both populations use the same uniform step-size quantizer and $e_p(j, k) = e_o(j, k)$ for all coefficients. Thus, (69) is satisfied if the practical coding strategy assigns each coefficient to the population that minimizes the actual number of bits $r_p(j, k)$ used by the fixed coding syntax. Under the coding syntax, a coefficient in band j falling in interval $[(i - \frac{1}{2})a, (i + \frac{1}{2})a]$, and assigned to the edge population uses

$$j + 1 - \log P_{e,j}(i) \text{ bits, where } P_{e,j}(i) = \int_{(i-1/2)a}^{(i+1/2)a} \frac{1}{\sqrt{2\pi\sigma_{e,j}^2}} e^{-x^2/(2\sigma_{e,j}^2)}\, dx$$

while the same coefficient assigned to the smooth population uses

$$-\log P_{s,j}(i) \text{ bits, where } P_{s,j}(i) = \int_{(i-1/2)a}^{(i+1/2)a} \frac{1}{\sqrt{2\pi\sigma_{s,j}^2}} e^{-x^2/(2\sigma_{s,j}^2)}\, dx$$

and $\sigma_{e,j}^2$ and $\sigma_{s,j}^2$ are the bounding variances for the two populations in band j. Thus, the required number of bits is minimized by assigning each coefficient to the edge population only if

$$j+1-\log P_{e,j}(i) < -\log P_{s,j}(i).$$

Noting that the expressions $-\log P_{*,j}(i)$ are the code lengths used by the two populations for representing the ith quantization interval, this condition can be simply stated: assign each coefficient to the edge population if and only if the ith code length for the edge population is shorter than the ith code length for the smooth population by at least $j+1$ bits. This strategy operates directly on coefficient values, and does not require an oracle for identifying either the number or location of coefficients in each population.

Case 2: $J_s < j \leq J_e$: For these bands, coefficients in the smooth populations are not coded, while coefficients in the edge population are quantized with a uniform-step-size quantizer, and entropy coded. Our practical strategy defines three modifications to the oracle's encoding strategy, each reducing the left-hand side of (69) for every coefficient coded, and together eliminating the need for the oracle.

First, the $i=0$ codeword of the edge population quantizer is replaced by a codeword with codelength zero, reflecting the fact that assigning a coefficient the $i=0$ symbol can be equivalently implemented by mapping the coefficient to the smooth population. Since this modification does not change any errors, but lowers the number of bits, inequality (69) is satisfied for each coefficient.

Second, the encoding strategy for edge coefficients is adapted to minimize an entropy-constrained distance criterion, given the fixed coding syntax (symbols and code lengths). Instead of following the oracle strategy of assigning to coefficient $c_{j,k}$ the symbol i minimizing

$$[c_{j,k}-ia]^2$$

the practical strategy selects i to minimize

$$[c_{j,k}-ia]^2-\lambda_a l_{e,j}(i)$$

where $l_{e,j}(i)$ denotes the code length (including addressing bits) used for assigning the ith symbol to an edge coefficient in band j. Given the first modification, these code lengths can be written

$$l_{e,j}(i)=\begin{cases}0, & \text{if } i=0\\ j+1-\log P_{e,j}(i), & \text{otherwise.}\end{cases}$$

Note that this practical strategy insures that inequality (69) is satisfied for every edge coefficient by minimizing the left-hand side of (69) over all possible encoding strategies. Intuitively, this strategy assigns something different than the minimum-distortion symbol only if the competing symbol offers a saving in code length that makes up for its higher distortion. The slope λ_a defines the worth in distortion of a saving in bits. The resulting scalar quantizer has uniformly spaced codewords, but *uses decision regions that are not uniformly structured*.

The third and final modification defining the practical encoder calls for the edge-population quantizer and encoding strategy to be applied to the coefficients of the smooth population. Given the first modification of the quantizer, this differs from the oracle-based strategy only when a coefficient from the smooth population is assigned a symbol $i \neq 0$. Since this case occurs only if

$$[c_{j,k}-ia]^2-\lambda_a l_{e,j}(i) < c_{j,k}^2$$

inequality (69) is satisfied for all smooth coefficients.

Since each of the three modifications monotonically lowers the left-hand side of (69), the overall encoding strategy satisfies (69) for every wavelet coefficient. It is also clear that after the modifications, since the same quantizer and encoding strategy is applied to both populations, the oracle is no longer needed for identifying the number and location of coefficients in each population. It is interesting to note that, although the mixture-model coding syntax and encoding strategy are originally designed to optimally match quantizers to two different populations of coefficients, the final practical coder can be described as applying a single quantizer, coupled with suitable entropy coding, to all coefficients of each wavelet band.

As mentioned at the end of the last section, the number of bits resulting from this coding strategy depends on the complexity of the input realization. Since the encoding strategy of this section reduces to the application of scalar quantizers to the coefficients of each band, the actual number of bits used to code a given input can be adjusted through the step-size parameter a, or equivalently λ_a.

C. Encoding General Processes

The piecewise-smooth model is a typical example of a "mixture-type process," since we can clearly classify its coefficients into two categories, depending on whether or not the associated wavelet overlaps a jump. In order to understand the performance of a mixture-type coder on more general processes, we somehow need to make a similar classification into two populations of slowly and rapidly decaying coefficients, even if this distinction does not appear as obviously as in the case of the PSM.

The more general processes that we now want to consider are precisely those which can be approximated at a prescribed rate N^{-2r} in the mean-square sense by an N-term combination of wavelets. According to the results on nonlinear approximation in Section II, an alternate description of such processes $f(t)=f(t,\omega)$ is given by the property

$$E(\|f\|_{B_{p,p}^{r,w}}^2) < \infty \tag{70}$$

with $1/p=1/2+r$ in the 1-D case. However, the smoothness property (70) turns out to be too weak for coding purposes: the cost of addressing the location of the large wavelet coefficients is unbounded for such processes since such coefficients can be located at arbitrarily high scales. This fact can be related to the lack of compactness in the embedding of $B_{p,p}^{r,w}$ into $\mathcal{L}^2$, i.e., the impossibility of covering the unit ball of $B_{p,p}^{r,w}$ with a controlled number of $\mathcal{L}^2$ balls of arbitrarily small radius.

We thus add the constraint that the process f is bounded by a fixed constant (say 1), yielding the "worst case estimate"

$$E(|\langle f,\psi_{j,k}\rangle|^2) \leq C2^{-j} \tag{71}$$

analogous to the slowly decaying population in the PSM. This bound restricts the location of large wavelet coefficients suffi-

ciently to allow them to be addressed with a finite cost. We now analyze how our mixture-model coder, motivated for coding the PSM, performs in coding this more general process. Our analysis follows two stages similar to our earlier analysis of the PSM process. We first assume an oracle that uses addresses to identify populations, and codes each population with an appropriate scalar quantizer and entropy coder. Later, based on entropy-constrained optimization principles, we define a practical coder without explicitly identifying populations.

We begin by noting from (12) that the Besov characterization (70) allows us to define a rearrangement of the wavelet coefficients in decreasing order of absolute value $(\tilde{c}_n)_{n>0}$ satisfying

$$E(|\tilde{c}_n|^2) \le Cn^{-(2r+1)}. \tag{72}$$

For this process, the oracle defines population 2 (analogous to the edge-population of the PSM) as the first N variables of $\{\tilde{c}_n\}_{n>0}$, and population 1 consists of all other coefficients. Addresses identify population 2 the same way as before: in band j, a $j+1$-bit address specifies the location of each coefficient in population 2, followed by a $j+1$ bit "end-of-address" symbol. For this process, there is no analog to the "smooth" population of the PSM: i.e., a large population of wavelet coefficients known to reside in certain bands, and thus not needing addressing information. Consequently, the oracle sets $J_s = 0$, using no bits for population 1 and setting all its coefficients to zero. Similar to the case of the PSM process, coefficients of population 2 are coded with a uniform-step-size scalar quantizer (step size a). Quantizer indexes are coded with an entropy code matched to a Gaussian random variable with variance equal to the bounding variance given in (71). Consequently, if $l_j(i)$ denotes the number of bits (including addressing) used to represent a coefficient in band j quantized to level i

$$l_j(i) = j + 1 - \log P_j(i), \quad \text{where}$$
$$P_j(i) = \frac{1}{\sqrt{2\pi C 2^{-j}}} \int_{(i-1/2)a}^{(i+1/2)a} e^{-x^2/(2C2^{-j})}\, dx. \tag{73}$$

The value J_e, denoting the highest band to be coded, is set so that, using (71), all coefficients in bands $j > J_e$ fall into the $i = 0$ bin of a uniform quantizer, thus defining the same J_e as in the PSM (64)

$$C2^{-J_e} = a^2/4 = \hat{D}_q. \tag{74}$$

Finally, N is selected to include in population 2 all reordered coefficients $\tilde{c}_n$ with variance exceeding $\hat{D}_q$. Applying (72) gives

$$N = C\hat{D}_q^{-1/(1+2r)} = C2^{J_e/(1+2r)}. \tag{75}$$

Like for the PSM process, we establish the rate-distortion performance of the mixture-model coder by independently bounding the distortion and the number of bits used in coding the bounded Besov process. Following the distortion analysis of the PSM smooth population, distortion is bounded by

$$D \le CN^{-2r} + N\hat{D}_q \tag{76}$$

with the two terms bounding distortion due to uncoded and coded coefficients, respectively. Using (75) to eliminate $\hat{D}_q$ gives

$$D \le CN^{-2r} \tag{77}$$

for some $C > 0$.

To bound the number of bits, we characterize the maximum number of bits needed to code each $\tilde{c}_n$ subject to (70) and (71). Note that overhead bits used for coding parameters and the "end-of-address" symbol for each band contribute no more than $C(\log N)^2$ bits, and are thus ignored in the analysis. Assume $\tilde{c}_n$ is in the jth band. If

$$P(\tilde{c}_n, i) = \text{Prob}[(i-1/2)a < \tilde{c}_n < (i+1/2)a]$$

then the number of bits needed to code $\tilde{c}_n$ is given by

$$R(\tilde{c}_n) = \sum_{i=-\infty}^{\infty} P(\tilde{c}_n, i) l_j(i) \tag{78}$$

with $l_j(i)$ defined in (73). $R(\tilde{c}_n)$ is maximized by finding the worst case probability distribution $P(\tilde{c}_n, i)$ subject to (71) and (72). Note that

$$\begin{aligned} l_j(i) &= j + 1 + \frac{1}{2}\log(2\pi C 2^{-j}) \\ &\quad - \log\left(\int_{(i-1/2)a}^{(i+1/2)a} e^{-x^2/(2C2^{-j})}\, dx\right) \\ &= C' + \frac{j}{2} + \frac{x_i^2}{2C2^{-j}} - \log a \end{aligned}$$

where we have again utilized the mean value theorem $((i-1/2)a \le x_i \le (i+1/2)a)$. Similar to arguments utilized in deriving (49), we can note that if $x \in [(i-1/2)a,\ (i+1/2)a]$ then $3x \ge x_i$ whenever $i \ne 0$, and with the given

$$a^2/4 = \hat{D}_q = C2^{-J_e} \le C2^{-j}$$

we can bound the contribution of the third term of the $l_j(i)$ expression to the summation in (78). We thus find that

$$C_1 + \frac{j}{2} \le R(\tilde{c}_n) \le C_2 + \frac{j}{2} + C_3 J_e \tag{79}$$

for constants C_1, C_2, and C_3. We conclude that these bounds on $R(\tilde{c}_n)$ are maximized when all $\tilde{c}_n$ fall into the highest allowed band J_e.

Given this characterization of the worst case Besov-bounded process, accounting for the number of bits is straightforward

$$R \le CNJ_e \tag{80}$$

where $C > 0$ is a constant. Using (75) to eliminate J_e yields

$$R \le C'' N \log N. \tag{81}$$

Together with (77) this gives

$$D(R) \le CR^{-2r}(\log R)^{2r} \tag{82}$$

for some constant C.

As a final stage in applying the mixture-model coder to the Besov-bounded process, we modify the encoding strategy as in

Section IV-B to eliminate the oracle in the encoding process. Although the oracle in this case only performs a reordering which could be performed by a practical coder, it is worth noting that the same performance can be achieved without reordering by using a single scalar quantizer applied to both populations in each band. The analysis here is identical to that in Section IV-B, so we briefly outline the three steps. First, the $i = 0$ quantization step is removed from each quantizer, since it can be implemented with zero bits by assigning the coefficient to population 1. Second, the standard minimum distortion quantization criterion is replaced with a minimum rate-distortion criterion in applying each of the quantizers. Finally, the modified quantizers are applied to both populations in each band. As shown in Section IV-B, each of these steps ensure that the resulting distortion and number of required bits continue to satisfy (82). Since the same quantizer is applied to both populations after the last modification, the encoding strategy does not need to either order the coefficients or to determine the correct value of N. Though the oracle-based coder is defined to encode exactly N coefficients, after the modifications the number of coded coefficients depends on the actual number of coefficients of each realization of the process that exceed the corresponding thresholds in each band.

V. Tree-Structured Approximation and Coding

In this section, we turn to a different coding strategy. It is related to a different family of "real-life" coders that exploit the efficiency of tree structures for identifying the significant coefficients in images. We demonstrate this efficiency in two classes of processes: those with functions from Besov spaces that guarantee a certain decay of large coefficients through scale, and those that are smooth except in some isolated regions. In the latter case, the multiscale decomposition is not only sparse but also exhibits a particular structure: as the scale j increases, the important coefficients tend to concentrate near the singularities, aligning themselves to a "tree-structure": if a coefficient $c_{j,k}$ in the wavelet expansion of the signal is numerically significant, it increases the probability that $c_{j',k'}$ is also numerically significant when $2^{-j'}k'$ is close to $2^{-j}k$. This particular structure can play an important role in coding, and we shall show that, both for the Besov spaces and the PSM model, tree structures provide another technique that can remove the logarithmic factors observed in Section III.

A. Tree-Structured Approximation

We shall now address nonlinear approximation processes where the tree structure is pre-imposed on the preserved coefficients. As before, we consider functions defined on the interval $[0, 1]$ and expanded into a wavelet decomposition

$$f = \sum_{j \geq 0} \sum_{k=0}^{2^j - 1} c_{j,k} \psi_{j,k} \tag{83}$$

where we have omitted the scaling function contribution for notational simplicity. By definition, a *tree* is a finite set T of indexes (j, k), $j \geq 0$, $k \in \{0, \ldots, 2^j - 1\}$, such that $(j, k) \in T$ implies $(j - 1, \lfloor k/2 \rfloor) \in T$, i.e., all "ancestors" of the point (j, k) in the dyadic grid also belong to the tree.

One can then consider the "best tree-structured approximation of f," by trying to minimize

$$\varepsilon_N = \left\| f - \sum_{(j,k) \in T} c_{j,k} \psi_{j,k} \right\|^2$$

over all trees T of cardinality N and all choices of $c_{j,k}$. However, the procedure of selecting the optimal tree is costly in computational time, in comparison to the simple reordering procedure that is used in (8).

A more reasonable approach is to use suboptimal tree selection algorithms. We now describe two of these algorithms.

Algorithm 1: Fix a threshold ε and define

$$E_\varepsilon = \{(j, k); |c_{j,k}| \geq \varepsilon\}.$$

Then define T_ε to be the smallest tree containing E_ε, i.e., T_ε contains all indexes (j, k) in E_ε and their "ancestors" (j', k'), $j' < j$, $k' = \lfloor 2^{j'-j} k \rfloor$.

Algorithm 2: Start from the initial tree $T_0 = \{(0, 0)\}$ and let it "grow" by the following procedure: given a tree T_N, define its "leaves" $\mathcal{L}(T_N)$ as the indexes $(j, k) \notin T_N$ such that $(j - 1, \lfloor k/2 \rfloor) \in T_N$. For $(j, k) \in \mathcal{L}(T_N)$, define the residual

$$r_{j,k} = \left(\sum_{I_{l,m} \subset I_{j,k}} |c_{l,m}|^2 \right)^{1/2} \tag{84}$$

with $I_{j,k} = [2^{-j}k, 2^{-j}(k + 1)]$. Choose $(j_0, k_0) \in \mathcal{L}(T_N)$ such that $r_{j_0,k_0} = \max_{(j,k) \in \mathcal{L}(T_N)} r_{j,k}$ and define $T_{N+1} = T_N \cup \{(j_0, k_0)\}$.

Note that the second algorithm can either be controlled by the cardinality N of the tree or by the size of the residual, i.e., by defining T_ε to be the smallest tree for which all residuals $r_{j,k}$, $(j, k) \in \mathcal{L}(T_N)$ are less than ε. Note also that T_N can also be defined as the set of indexes (j, k) corresponding to the $N + 1$ largest values of $r_{j,k}$.

We are now interested in describing those functions such that the error of approximation ε_N produced by such algorithms behaves like N^{-2r}. Another way of stating that ε_N behaves like N^{-2r} is by saying that we have a control on the cardinality of T_ε according to

$$\#(T_\varepsilon) \leq C \varepsilon^{-p} \tag{85}$$

with p such that $1/p = 1/2 + r$, and $C > 0$ is a constant. Indeed, using the fact that for both algorithms we have $|c_{j,k}| \leq \varepsilon$ if $(j, k) \notin T_\varepsilon$, we can derive from (85) the error estimate

$$\begin{aligned}
\left\| f - \sum_{(j,k) \in T_\varepsilon} c_{j,k} \psi_{j,k} \right\|^2 &= \sum_{(j,k) \notin T_\varepsilon} |c_{j,k}|^2 \\
&= \sum_{n > 0} \sum_{(j,k) \in T_{2^{-n}\varepsilon} \setminus T_{2^{-n+1}\varepsilon}} |c_{j,k}|^2 \\
&\leq C_1 \sum_{n > 0} 2^{-2n} \varepsilon^2 \#(T_{2^{-n}\varepsilon}) \\
&\leq C_2 \varepsilon^{2-p} \left[\sum_{n > 0} 2^{-(2-p)n} \right]
\end{aligned}$$

for constants C_1 and C_2, i.e., with a $C > 0$ we have

$$\left\| f - \sum_{(j,k)\in T_\varepsilon} c_{j,k}\psi_{j,k} \right\|^2 \le C\varepsilon^{2-p}. \tag{86}$$

Combining again with (85), we obtain

$$\left\| f - \sum_{(j,k)\in T_\varepsilon} c_{j,k}\psi_{j,k} \right\|^2 \le C\#(T_\varepsilon)^{-2r} \tag{87}$$

for some $C > 0$.

So far, the approximation properties of both algorithms are not fully understood in that we do not know exactly the smoothness space describing those functions such that (85) holds. The following suboptimal result shows that this space is nearly as large as the weak Besov space $B^{r,w}_{p,p}$, $1/p = 1/2 + r$, which we encountered in the context of standard thresholding.

Proposition 8: Assume that the function f belongs to $B^r_{q,\infty}$, with $1/q < 1/2 + r$, and let $B(f) = \|f\|_{B^r_{q,\infty}}$. Then, both algorithms satisfy (85) and have the approximation property

$$\|f - f_N\| \le CB(f)N^{-r} \tag{88}$$

where the tree-structured approximation f_N is obtained in the first algorithm by choosing the smallest ε such that $\#(T_\varepsilon) \le N$, or by N steps of the second algorithm.

Proof: We start from the characterization of $B^r_{q,\infty}$ from wavelet coefficients (see (A19) in the Appendix on Besov spaces): a function f is in $B^r_{q,\infty}([0, 1])$ if and only if we have the estimate

$$\sum_k |c_{j,k}|^q \le C[B(f)]^q 2^{-(qr+q/2-1)j} \tag{89}$$

independent of j. Since $(qr + q/2 - 1) > 0$, (89) is equivalent to

$$\sum_{l\ge j}\sum_k |c_{l,k}|^q \le C[B(f)]^q 2^{-(qr+q/2-1)j}. \tag{90}$$

Now observe that since $q \le 2$, we have

$$|r_{j,k}|^q = \left[\sum_{I_{l,m}\subset I_{j,k}} |c_{l,m}|^2\right]^{q/2} \le \sum_{I_{l,m}\subset I_{j,k}} |c_{l,m}|^q$$

so that (90) implies

$$\sum_k |r_{j,k}|^q \le C[B(f)]^q 2^{-(qr+q/2-1)j}. \tag{91}$$

We start by discussing the first algorithm. We denote by $N_j(\varepsilon)$ the number of coefficients at scale j that satisfy $|c_{j,k}| \ge \varepsilon$. From (89), it follows that

$$N_j(\varepsilon) \le C\varepsilon^{-q}[B(f)]^q 2^{-(qr+q/2-1)j}. \tag{92}$$

Note that for j large enough, the right-hand side becomes less than 1 so that there are no more coefficients above the threshold. On the other hand, we also have $N_j(\varepsilon) \le 2^j$. We denote by j_ε the scale such that $2^{j_\varepsilon} \approx C\varepsilon^{-q}[B(f)]^q 2^{-(qr+q/2-1)j_\varepsilon}$, i.e.,

$$2^{j_\varepsilon} \approx C'[B(f)]^{1/(r+1/2)}\varepsilon^{-1/(r+1/2)}, \qquad \text{for some } C' > 0.$$

We now embed the tree T_ε into a larger tree $\tilde{T}_\varepsilon$ composed of

- all indexes (j, k) for $j < j_\varepsilon$;
- all indexes (j, k) with $j \ge j_\varepsilon$ for which $|c_{j,k}| > \epsilon$, as well as their "ancestors" $(j', \lfloor 2^{j'-j}k \rfloor)$ for $j_\varepsilon \le j' < j$.

Using this decomposition together with (91), we estimate the cardinality of T_ε as follows:

$$\begin{aligned}\#(T_\varepsilon) &\le 2^{j_\varepsilon} + \sum_{j\ge j_\varepsilon}(j - j_\varepsilon)N_j(\varepsilon)\\ &\le C_1 2^{j_\varepsilon}\left[1 + \sum_{j\ge j_\varepsilon}(j - j_\varepsilon)2^{-(qr+q/2-1)(j-j_\varepsilon)}\right]\\ &\le C_2 2^{j_\varepsilon} \le C_3[B(f)]^{1/(r+1/2)}\varepsilon^{-1/(r+1/2)}\\ &= C[B(f)]^p\varepsilon^{-p}\end{aligned} \tag{93}$$

which is (85), for constants C_1, C_2, C_3, and C.

We now turn to the second algorithm. Again, defining $M_j(\varepsilon)$ as the number of coefficients at scale j that satisfy $r_{j,k} \ge \varepsilon$, it follows from (91) that

$$M_j(\varepsilon) \le C[B(f)]^q\varepsilon^{-q}2^{-(qr+q/2-1)j} \tag{94}$$

and we also have $M_j(\varepsilon) \le 2^j$. With the same definition of j_ε as for the first algorithm, we estimate

$$\begin{aligned}\#(T_\varepsilon) &\le \sum_{j\ge 0} M_j(\varepsilon)\\ &\le \sum_{j<j_\varepsilon} 2^j + C_1\sum_{j\ge j_\varepsilon}[B(f)]^q\varepsilon^{-q}2^{-(qr+q/2-1)j}\\ &\le C_2 2^{j_\varepsilon} \le C_3[B(f)]^{1/(r+1/2)}\varepsilon^{-1/(r+1/2)}\\ &= C[B(f)]^p\varepsilon^{-p}\end{aligned} \tag{95}$$

which is (85), for constants C_1, C_2, C_3, and C.

We already saw that (85) implies the error estimate (87). Here, since the constants in (93) and (95) contain $[B(f)]^p$, we obtain for both algorithms

$$\left\| f - \sum_{(j,k)\in T_\varepsilon} c_{j,k}\psi_{j,k} \right\|^2 \le C[B(f)]^p\varepsilon^{2-p}. \tag{96}$$

For a given N, choosing the smallest ε so that $\#(T_\varepsilon) \le N$ and combining the estimates on $\#(T_\varepsilon)$ with (96), we thus have for both algorithms the desired estimate

$$\|f - f_N\|^2 \le C[B(f)]^2 N^{-2r}$$

where f_N is the corresponding approximation. □

The result that we have proved shows that, for functions that have r derivatives in $\mathcal{L}^q$, with $1/q < 1/2 + r$, the rate N^{-2r} can be achieved while imposing the tree structure. Note that this property is barely more restrictive than $f \in B^r_{p,p}$, $1/p = 1/2 + r$, related to the nonlinear approximation in Proposition 2, since q is allowed to be arbitrarily close to p.

In the stochastic framework, the PSM is well adapted to tree-structured approximation. Note that according to the above result, we would obtain an immediate result of tree-structured approximation with both algorithms for the PSM if we could prove that $E(\|f\|^2_{B^r_{q,\infty}}) < \infty$, for some q such that $1/q < 1/2 + r$.

However, this property does not seem to hold. Nevertheless, we shall prove an approximation result in the mean-square sense, in the context of the first algorithm (similar results can be proved for the second one).

Proposition 9: For the piecewise-stationary model, if we consider for each realization the tree T_ε generated by the first algorithm, we have

$$E\left(\left\|f-\sum_{(j,k)\in T_\varepsilon} c_{j,k}\psi_{j,k}\right\|^2\right)\le C\varepsilon^{2-p} \tag{97}$$

and

$$E(\#(T_\varepsilon))\le C\varepsilon^{-p} \tag{98}$$

where p is such that $1/p=1/2+r$. Thus, if N is the expected number of coefficients and ε_N the mean-square error, we have

$$\varepsilon_N\le CN^{-2r}. \tag{99}$$

Proof: We place ourselves in the event where the number of discontinuities equals L, and we split each trajectory $f=\sum_{j,k}c_{j,k}\psi_{j,k}$ into two parts:

$$f_1=\sum_{(j,k)\in Y} c_{j,k}\psi_{j,k}$$

and

$$f_2=\sum_{(j,k)\notin Y} c_{j,k}\psi_{j,k}$$

where Y is the set of indexes (j,k) such that a discontinuity d_n is contained in the support of $\psi_{j,k}$ (note that Y changes for each realization of f).

Using the estimates on linear approximation of Proposition 2, and the $\mathcal{C}^{2r}$ smoothness of the local autocorrelation function $\mathcal{R}$, we know that

$$E\left(\sum_k |\langle f_2,\psi_{j,k}\rangle|^2\right)\le C2^{-2rj}. \tag{100}$$

For the function f_1, since f is bounded, we have the crude estimate

$$|c_{j,k}|\le C2^{-j/2}. \tag{101}$$

Using these estimates, we can compute the expected cardinality of T_ε from the expected number $E(N(j,\varepsilon))$ of coefficients at level j such that $|c_{j,k}|\ge\varepsilon$: we have indeed

$$E(\#(T_\varepsilon))\le\sum_{j\ge0} jE(N(j,\varepsilon)).$$

For the f_1 part the contribution to $E(N(j,\varepsilon))$ is null for values of j such that $C2^{-j/2}\le\varepsilon$ and always less than $L\times\mathrm{supp}\,(\psi)$ for smaller values of j, where $\mathrm{supp}\,(\psi)$ is the width of the compact mother wavelet support (see also footnote 4 in Section III-B). For the f_2 part, we derive from (100) that the contribution to $E(N(j,\varepsilon))$ is less than $C2^{-2rj}\varepsilon^{-2}$, while we know that it is also less than 2^j. Thus, the two contributions give us

$$E(\#(T_\varepsilon))\le\sum_{j\ge0} j\min\{2^j, C2^{-2rj}\varepsilon^{-2}\}+CL[\log(\varepsilon)]^2. \tag{102}$$

The summation in the first term (corresponding to f_1) should be divided into $j\le j_\varepsilon$ such that $2^j\le C2^{-2rj}\varepsilon^{-2}$ and $j\ge j_\varepsilon$ such that $2^j\ge C2^{-2rj}\varepsilon^{-2}$. Both are dominated by $C2^{j_\varepsilon}=C\varepsilon^{-2/(1+2r)}=C\varepsilon^{-p}$. We thus obtain

$$E(\#(T_\varepsilon))\le C\varepsilon^{-p}+CL[\log(\varepsilon)]^2. \tag{103}$$

Averaging over all possible values of L, we see that the second term (corresponding to f_2) remains negligible so that we finally obtain

$$E(\#(T_\varepsilon))\le C\varepsilon^{-p} \tag{104}$$

which is (98).

For the error estimate, we write, similar to the deterministic case

$$\begin{aligned}
&E\left(\|f-\sum_{(j,k)\in T_\varepsilon} c_{j,k}\psi_{j,k}\|^2\right)\\
&\quad=E\left(\sum_{(j,k)\notin T_\varepsilon}|c_{j,k}|^2\right)\\
&\quad=\sum_{j\ge0}E\left(\sum_{(j,k)\in T_{2^{-j-1}\varepsilon}\setminus T_{2^{-j}\varepsilon}}|c_{j,k}|^2\right)\\
&\quad\le\sum_{j\ge0}(2^{-j}\varepsilon)^2E(\#(T_{2^{-j-1}\varepsilon}))\\
&\quad\le C\sum_{j\ge0}(2^{-j}\varepsilon)^{2-p}\\
&\quad=C\varepsilon^{2-p}\sum_{j\ge0}2^{-j(2-p)}=C'\varepsilon^{2-p}
\end{aligned}$$

which is (97). Combining both, we obtain (99). □

B. Tree-Structured Coding

We shall now derive a compression result from the tree-structured approximation properties that we have introduced above. Like the mixture approach in Section IV, the tree structure will allow us to remove the logarithmic factor that was observed in Section III, and to obtain a result in a rate-distortion sense over the Besov classes and for the piecewise-stationary model. More specifically, we shall construct a coder based on successive tree-structure approximation of our signal f.

We shall build a coder that achieves $D(R)\le CR^{-2r}$, whenever a similar rate is produced by nonlinear tree-structured approximation. For a given $\varepsilon>0$, we operate the encoding in three steps.

Step 1: Tree Analysis of the Signal: Define for $l=0,1,\ldots,$ the trees $T_{2^l\varepsilon}$ produced by one of the two algorithms. We then define the "layers" $\Delta_l=T_{2^l\varepsilon}\setminus T_{2^{l+1}\varepsilon}$.

Step 2: Quantization of the Coefficients: We only encode the coefficients with indexes in T_ε. We allocate l bits for each coefficient with index $(j,k)\in\Delta_l$. The total number of bits for quantizing these coefficients can thus be estimated as follows:

$$R_{\mathrm{quant}}=\sum_{l\ge0} l\#(\Delta_l)\le\sum_{l\ge0} l\#(T_{2^l\varepsilon}). \tag{105}$$

The total distortion is controlled by the truncation error $\|f-\sum_{(j,k)\in T_\varepsilon}c_{j,k}\psi_{j,k}\|^2$ and the quantization error. Since we are using l bits for each coefficient with index $(j,k)\in\Delta_l$,

the contribution to this second error of every single coefficient is less than ε. The global quantization error is thus estimated by $\varepsilon^2 \#(T_\varepsilon)$.

Step 3: Quantization of the Trees: We also need to encode the sets Δ_l, $l \geq 0$, or, equivalently, the sets $T_{2^l\varepsilon}$. Clearly, a tree T of cardinality N can be encoded in $\mathcal{O}(N)$ bits, since at each scale j it is sufficient to put twice as many bits as the number of elements in the tree at the previous scale $j-1$, in order to indicate if an index is contained in the tree $(b=1)$ or not $(b=0)$. Consequently, the total number of bits needed to encode the sets Δ_l can be estimated by

$$R_{\text{addr}} \leq \sum_{l \geq 0} \#(T_{2^l\varepsilon}). \tag{106}$$

We shall now prove that the performance of this coder over a deterministic class can be derived from the performance of tree-structure coding: given $r>0$ and p such that $1/p=1/2+r$, we consider a class B such that

$$\sup_{f \in B} \#(T_\varepsilon f) \leq C\varepsilon^{-p} \tag{107}$$

where $T_\varepsilon f$ is the tree associated to the function f and the parameter ε.

From this estimate, we obtain that the quantization error is estimated by $C\varepsilon^{2-p}$, and so is the truncation error according to (86). We thus have $D \leq C\varepsilon^{2-p}$. Finally, according to (105) and (106), the number of bits is bounded by

$$\begin{aligned} R &\leq R_{\text{quant}} + R_{\text{addr}} \\ &\leq \sum_{l \geq 0} l \#(T_{2^l\varepsilon}) + \sum_{l \geq 0} \#(T_{2^l\varepsilon}) \\ &\leq C \sum_{l \geq 0} (1+l)\varepsilon^{-p} 2^{-lp} \\ &\leq C\varepsilon^{-p}. \end{aligned}$$

Summarizing, we have proved the following result.

Proposition 10: If $D(R, f)$ is the distortion obtained by the previous coder applied on the function f with R bits at our disposal, we have

$$\sup_{f \in B} D(R, f) \leq CR^{-2r}. \tag{108}$$

In particular, we can apply this result to the Besov classes $B = B^r_{q,\infty}$, $1/q < 1/2 + r$, according to Proposition 8. For these classes, the exponent in (108) is actually optimal in the sense that it achieves the theoretical entropy bounds (see [6]).

We can also rephrase this result for the PSM model, using the estimate (98) in place of (107). More generally, we have the following.

Proposition 11: Let f be a stochastic process such that

$$E(\#(T_\varepsilon)) \leq C\varepsilon^{-p}. \tag{109}$$

Then, the previous coder applied to f gives

$$D \leq CR^{-2r} \tag{110}$$

where D is the mean-square error and R the average number of bits.

Of course, according to Proposition 9, this applies to the piecewise-stationary stochastic process.

VI. Generalizations, Shortcomings, and Conclusion

This paper has analyzed the efficiency of coding certain classes of random processes with wavelet-based strategies derived from nonlinear approximation. We have shown that the asymptotic coding efficiency of these algorithms parallel the asymptotic approximation accuracy achieved by nonlinear approximation, i.e., the relation between distortion and number of bits used in coding follows the same power law as the relationship between distortion and approximation order. This section reviews and interprets these results and discusses their relation to the design and performance of current state-of-the-art algorithms for coding images.

We begin by remarking that most of the results in this paper, although stated in the 1-D setting for the sake of simplicity, have straightforward generalization to higher dimensions: the relation between the order of smoothness and the decay rate of linear/nonlinear approximation error changes (see Appendix), but the same coding strategies, generalized naturally to higher dimensions, allow us to derive similar asymptotic behavior for compression and approximation algorithms.

For all the processes we have considered in our analysis, smoothness of order r guarantees that the magnitudes of the wavelet coefficients, ordered in decreasing order, decay sufficiently fast to allow approximation error to be bounded by (13). Although this decay is common to the spaces $B^r_{2,\infty}$, $B^{r,w}_{p,p}$, $B^r_{p,\infty}$, and the PSM model considered in various sections of our analysis, these processes differ in the nature of rearrangement required to uncover the desired decay. For $B^r_{2,\infty}$, no such reordering is necessary, and linear approximation, analyzed in Section III-A, achieves coding efficiency reflecting the underlying smoothness r. However, for all the processes considered later, linear approximation has very poor coding efficiency because, by coding coefficients in their natural order, it fails to exploit the underlying decay properties of the process. All the coding strategies studied in later sections exploit this underlying decay rate by either explicitly or implicitly encoding some ordering information about the coefficients.

For the PSM defined at the end of Section II, the coefficients are characterized explicitly as a mixture of two classes of coefficients: smooth coefficients decaying like $n^{-(2r+1)}$ in the mean square, and edge coefficients decaying like n^{-1}. The idealized coder analyzed in Section III-B explicitly encodes this classification by transmitting the number and location of edges. Though this does not specify the exact ordering of coefficient magnitudes, it gives enough information to rearrange the coefficients into a sequence satisfying the desired decay rate. Sound quantization strategies applied to this rearranged sequence achieves rate-distortion performance reflecting the smoothness r. The practical coder developed in Section IV takes a more general approach to classifying coefficients, not so explicitly tied to the structure of the PSM model. Instead of assuming a specific structure created by edges, the practical coder provides addresses for assigning an arbitrary set of coefficients in each band to one class (all others are assigned to a second class). While this is a less efficient way to rearrange coefficients of the PSM model, this more general coder still achieves the desired asymptotic rate-distortion performance.

In Section IV-C, the same practical coder developed for the PSM is applied to a process of bounded functions from the space $B^{r,w}_{p,p}$. Though such functions allow a very rich rearrangement of coefficients, constrained only by the size of coefficients that can be found in each band, it is shown that the general addressing scheme of the practical coder, based on classifying coefficients into two classes, allows it to achieve asymptotic rate-distortion performance reflecting the underlying smoothness r. (Note: the extra $(\log R)^r$ factor reflects the highly unstructured rearrangements that are needed to code this class of functions.)

Finally, Section V analyzes tree-structured techniques for coding processes of functions from $B^r_{q,\infty}$, with $1/q < 1/2 + r$. For this space of functions, (92) constrains the number of large coefficients that can be found at scale j, a constraint that did not apply in the previous section. The tree-structured coder classifies coefficient indexes into sets, denoted Δ_l in Section V-B, whose structures are specified via a collection of embedded trees. For any function from $B^r_{q,\infty}$, with $1/q < 1/2 + r$, the sets $\{\Delta_l\}$ can be viewed as specifying a reordering of the wavelet coefficients into a sequence with magnitudes bounded by $Cn^{-(r+1/2)}$. The tree-structured constraint allows the sets to be encoded with very few $(\mathcal{O}(N))$ bits. It is worth noting that these results use (92) to bound the number of large coefficients found at each scale j, but do not assume any structure for the location of large coefficients in any band. Thus, our analysis does not exploit any tree-structures in the location of coefficients, but rather shows that an arbitrary collection of large coefficients satisfying (92) can be "covered" by a tree with reasonable efficiency (i.e., the covering tree does not include *too* many other coefficients). For processes, like the PSM, for which spatial clustering causes large coefficients to be aligned along tree-like structures, the tree-structured coder offers even more coding advantages, but these advantages only improve constants and do not change the asymptotic decay characteristics of our analysis.

In relating our analysis to practical coding design issues, it is important to recognize the limitations of using asymptotic performance results to characterize coding of real images. Difficulties arise due to several factors.

- Discretization: our analysis considers real-valued functions defined on a continuous domain. Real images are represented as samples on a discrete grid, and taking on integer values in a finite range (typically $\{0, \ldots, 255\}$).
- Spatial mixing: our analysis considers functions from a single space of functions. Real images usually contain a rich mixture of spatial regions, characterized with perhaps many different smoothness parameters.
- Noise: even regions within real images that appear to be well modeled by some function space are almost always distorted by additive white noise.

Due to discretization, the original image can be specified losslessly with a finite number of bits describing a finite number of wavelet scales. While this forces a trivial asymptotic behavior after some high, but finite, number of bits (i.e., zero distortion is reached), even at lower bit rates the finite number of nonzero wavelet scales undermines the asymptotic performance of nonlinear approximation. It is well known that the rate-distortion behavior for coding any given coefficient (or fixed set of coefficients) follows an exponential $D(R)$ decay (i.e., $D(R) \leq Ce^{-KR}$). As R increases, coders allocate additional increments of bits for two purposes: i) to refine the quantization of the current set of "significant" coefficients, and ii) to add new coefficients to the set of significant coefficients. (Note: the term "significant" was first used in [22] to refer to those coefficients that are not automatically set to zero by the classification itself. In this paper, significant coefficients would be all those assigned addresses in Section IV, and all those contained in trees in Section V.) Since allocations of type i) produce exponential decay in $D(R)$, overall coding performance is dominated by the $D(R)$ decay resulting from allocations of type ii). The efficiency of nonlinear approximation comes from its ability to address large coefficients at both fine and coarse scales, allowing it to maximize the decay from type ii) allocations. If the original image contains only a finite number of nonzero coefficients, at some high enough bit rate all coefficients are classified as significant, and $D(R)$ decay becomes exponential. The polynomial decay in $D(R)$ predicted by our analysis will only be seen at bit rates low enough that the number of significant coefficients is a small fraction of the total number of coefficients.

For an image consisting of many regions drawn from different smoothness spaces, asymptotic performance is governed by the lowest smoothness parameter. However, at any given finite number of bits, the characteristics of the $D(R)$ curve are mostly influenced by the regions that receive the largest allocations of incremental increases in the total number of bits. Thus, at moderate bit rates, all regions contribute to the decay of the $D(R)$ curve, with relative importance of the regions governed by their energy and size.

Even ignoring discretization and mixture of regions, all real images are subject to additive noise. Strictly speaking, the asymptotic performance for coding a sum of signal plus noise is dominated by the noise only (assuming the smoothness of the noise is much less than that of the signal). However, if the noise energy is much less than the signal energy (usually the case), and if we consider coding performance at low to moderate bit rates, it is clear that the large-signal coefficients are responsible for the bit rate that is used, and the reduction in distortion. Thus, rate-distortion decay at such bit rates is dominated by the signal component and should reflect the signal smoothness. As bit rate increases, we expect a transition in decay characteristics to reflect the lower smoothness of the noise.

Combining all these factors, we conclude that the results of our asymptotic analysis are not so much important for how they predict very high bi trate performance, as for how they predict the $D(R)$ decay as finite-bit-rate allocations are made to the collection of regions in each image. Roughly speaking, our analysis suggests that ideal decay is achieved if some form of addressing successfully identifies the largest coefficients in each region to which bits should be allocated. It is reasonable to expect that a coder achieving $D(R)$ decay that fails to reflect the underlying smoothness of each region is at a significant disadvantage compared to one that achieves nearly ideal decay. This remark must be qualified by the observation that the constants ignored in our analysis also need to be considered. For a given coder, the

constant is effected by how many bits are used for addressing, how successfully the addressing finds the largest coefficients, and how efficiently the coder encodes the coefficients that are found. In today's state-of-the-art image coders, these issues are carefully balanced against each other to optimize coding efficiency. In the following paragraphs, we review several approaches used by today's practical wavelet-based coders, highlighting the role of nonlinear approximation in each. Performance results for most of these coders are available on the Internet (see [25]).

The idea of using addressing to identify some subset of "significant" coefficients is central to the contributions of the Shapiro coder [22]. Significant coefficients with respect to a threshold are those coefficients whose magnitudes exceed the threshold, and the coder identifies tree-structured groups of coefficients containing any significant coefficient. In order to generate an embedded bit stream, significance is identified at a diadically decreasing sequence of thresholds. Like the coder of Section V, the number of bits used for coding the significant coefficients is determined by the threshold at which they become significant. Though many detailed issues contribute to the good performance of the Shapiro coder, its use of tree-structured encoding of significance is certainly a major contributor, and the approach it uses is quite similar to that studied in Section V. Later refinements of embedded tree-structured approaches (e.g., the SPIHT algorithm of [21]) improve upon the performance of the Shapiro coder by putting probability models on the tree structures used to classify significance, making it more efficient to code certain structures of large coefficients than others. While these models match well the clustering of edge coefficients in the PSM, such refinements exploit structures that are completely ignored in the Besov-space models of functions. In fact, a typical function from any of the Besov spaces should be coded equally well by the SPIHT or the Shapiro algorithm. The fact that SPIHT shows a significant performance advantage coding natural images suggests that, although Besov spaces accurately model the scattering of significant coefficients through various scales of the wavelet transform, they fail to model important elements of structure in that scattering. SPIHT at least partially succeeds in exploiting that structure.

The SFQ algorithm [27] is another coder using a tree-structured approach for addressing significant coefficients. Unlike the two algorithms in the previous paragraph, the SFQ does not use bit-level coding but codes significant coefficients with uniform scalar quantization and entropy coding, very much like the approach of Section IV. Unlike Section IV, however, the addresses of the coefficients to be coded are encoded with a"significance map" that is constrained to have the structure of a pruned tree. Coefficients in the significance map are coded, and all others are set to zero. The SFQ uses an optimization procedure to prune the full tree of wavelet coefficients in order to minimize $D(R)$. The fact that SFQ is slightly more efficient than SPIHT for coding natural images shows that the advantages of tree-structured addressing can be realized with both bit-level coders and more standard coefficient quantizers.

Several top performing wavelet-based algorithms are more closely related to the coder of Section IV, using addressing mechanisms that are not tree-structured. In the SR (stack-run) algorithm of [24], the significance map of each band is encoded using a run-length algorithm (codes the length of runs of insignificant coefficients), and then the significant coefficients are entropy coded. Unlike the tree-structured algorithms, the SR algorithm uses no interband dependencies, coding each scale of the wavelet transform separately. The run-length coding can be recognized as a more efficient way to address significant coefficients than that used in Section IV. Instead of sending the distance from the first coefficient (i.e., a standard address), the run-lengths send the distance from the last significant coefficient. Besides the fact that the addresses are smaller, run-length coding, coupled with entropy coding, also partially exploits the statistical clustering of the significant coefficients in natural images.

Another coder related to the one in Section IV is the subband coder of [14]. Initially, it is not clear how this coder is related to nonlinear approximation. It applies a wavelet-like decomposition, and then codes all coefficients (i.e., no addressing is used) in each band with a form of entropy coding. The key feature of this coder is that two parameters are transmitted along with each band of coefficients and used to match a generalized Gaussian density (GGD) to the coefficients of the band. The GGD is given by

$$f(x) = \left[\frac{\nu\eta(\nu,\sigma)}{2\Gamma(1/\nu)}\right] \exp(-[\eta(\nu,\sigma)|x|]^{\nu}) \qquad (111)$$

where

$$\eta(\nu,\sigma) = \sigma^{-1}\left[\frac{\Gamma(3/\nu)}{\Gamma(1/\nu)}\right]^{1/2}. \qquad (112)$$

The parameters σ and ν allow independent control of the standard deviation and shape of the distribution, and are computed to best fit the coefficients in each band. Decreasing the value ν concentrates more probability near zero (e.g., for a Gaussian density $\nu = 2.0$ and for a Laplacian density $\nu = 1.0$). For the processes considered in our analysis, at increasingly fine scales the percentage of significant coefficients approaches zero, making the parameter ν approach zero, and resulting in a large concentration of probability near zero. Using the probabilities from this model, the coder in [14] uses a run-length code to identify runs of the zero-valued coefficients, and codes the significant coefficients with another entropy coder. Since runs of zero symbols are equivalent to addresses of nonzero symbols, we recognize an implicit role for addressing in this coder. This is similar to the coder of Section IV which, though initially developed with explicit addresses, reduced at the end of Section IV-B to an algorithm that applied uniform scalar quantization and a suitably designed entropy coder to all coefficients. It is not surprising to note that the performance results reported in [14] are not as good as other top wavelet-based coders. Though it exploits the structure of Besov space functions, this coder is based on a first-order probability model of fine-scale coefficients, and it is thus unable to exploit any structure in the position of significant coefficients in these scales, i.e., the number of bits used to code a spatial cluster of significant coefficients is the same as the number used to code the same coefficients scattered uniformly through

the band. This shortcoming, shared by the coder of Section IV, highlights a limitation of Besov-space modeling to images.

Two of the top performing image coders are also related to the coder of Section IV, though they both incorporate very sophisticated quantization techniques unrelated to our analysis. In the classification-based coder of [15], blocks of wavelet coefficients in each band are classified into four classes and coded with quantizers and entropy coders optimized for each class. Serving a similar role as addressing, the classification information is coded and transmitted as overhead. Block sizes are selected to optimize the tradeoff between overhead bits and classification accuracy. The classification itself is designed to optimize coding efficiency. Because this coder uses a variety of advanced modeling and quantization techniques, it is difficult to draw clear conclusions about the importance of nonlinear approximation. However, classification of coefficients into four classes allows more accurate reordering of the wavelet coefficients than the two-class mixture modeling of Section IV, and this feature makes a significant contribution to its very good performance results.

The estimation–quantization (EQ) coder of [18] models the coefficients at each scale of the wavelet transform as samples from an uncorrelated Gaussian process with spatially varying variance. The EQ decoder makes local estimates of the variance field based on previously decoded neighboring coefficients, and applies to each coefficient a quantizer and entropy code optimized for the local variance estimate. Although no explicit addressing information is coded, some small side-information is sent to help improve the accuracy of the variance estimates. This side-information and the estimates themselves play a similar role to addressing, since they effectively allow the coder to reorder wavelet coefficients for the purpose of optimal quantization. It is clear that the EQ coder very aggressively exploits features of images modeled by the PSM but not by the Besov spaces considered in our analysis. For a typical function from a Besov space, it should not be possible to accurately estimate variances of coefficients based on previously decoded neighboring coefficients. The very good coding performance of the EQ algorithm is evidence of structure in the distribution of coefficient energy within each wavelet band.

Because linear and nonlinear approximation methods are often associated with classical transform-based and wavelet-based coding algorithms, respectively, it is natural to ask how our analytical results relate to the performance of standard DCT-based image coding algorithms. The answer is not so clear. While Section III-A shows linear approximation to be very inefficient at coding the PSM model (the same would be true for the Besov-space models), the coder modeled in that analysis applies a global transform to the entire function, unlike standard block transform-based image coders which apply transforms to small blocks of pixels in the image. The inefficiency of linear approximation comes from the spread of localized energy (e.g., from impulses or edges) throughout all the coefficients of the global transform. But, in standard coders localized energy is at least confined to a finite number of block coefficients. To make any statements about the asymptotic efficiency of block-based transform coding, we would have to extrapolate how such a coder would operate on increasingly fine-resolution images. If the spatial support (measured in the continuous scene) is held constant as the resolution increases, the block size (measured in pixels) would grow, and asymptotic performance would reflect the inefficiency of linear approximation. On the other hand, if the block size (in pixels) stays constant, localized energy would remain confined to a finite set of coefficients. In this case, however, another layer of coding issues arises in order to exploit the correlation among blocks, and to efficiently address the high-energy blocks and large coefficients within blocks. In fact, standard image and video coding algorithms [16], [20] incorporate interblock prediction to efficiently code the DC coefficient, and run-length coding to efficiently address the large high-frequency coefficients. The coder of [26] is a block transform-based coder that directly applies techniques from nonlinear approximation to improve efficiency. This coder defines a tree structure on the DCT coefficients designed to efficiently address groups of coefficients corresponding to edges of various spatial orientation. By efficiently addressing these commonly occurring collections of significant coefficients, and by applying standard quantization to the significant coefficients, significant performance improvements are achieved over standard algorithms. We conclude that standard transform-based image coders fall somewhere in between our analysis of linear and nonlinear approximation. The block structure of the transform forces them to trade off the efficiency of large, low-frequency basis functions against the efficiency of localized, high-frequency basis coupled with efficient addressing. However, for coding sample images representing only a finite range of frequencies, experience shows that reasonable tradeoffs can be reached, allowing block transform-based coders to approach the performance of the best wavelet-based coders. But to achieve such performance, even these coders incorporate techniques motivated by nonlinear approximation.

Last but not least, it is important to keep in mind the differences between typical Besov-space functions and images. As mentioned earlier, general Besov-space functions have no structure beyond certain decay properties of the wavelet coefficients whereas images contain many structures that imply much more than these decay properties. It is clear that the further an image-compression algorithm exploits the delineating properties of images, the better its performance will be. The results and algorithms presented in this paper are valid (in an average or worst case sense) for the general objects one would encounter in Besov spaces mentioned in this paper. Our results should therefore not be viewed as the ultimate in image compression but more as general guidelines relevant to a broader space which future successful image compression algorithms should be aware of.

APPENDIX
A PRIMER ON BESOV SPACES

There exist many different ways of measuring the smoothness of a function f. The most natural one is certainly the order of differentiability, i.e., the maximal index m such that $f^{(m)} = (\frac{d}{dx})^m f$ is continuous. To this particular measure of smoothness, we can associate a class of *function spaces*: if I is

an interval of $\mathbb{R}$, we denote by $\mathcal{C}^m(I)$ the space of continuous functions which have bounded and continuous derivatives, up to the order m. This space can be equipped with the norm

$$\|f\|_{\mathcal{C}^m(I)} := \sup_{x\in I}|f(x)| + \sup_{x\in I}|f^{(m)}(x)| \tag{A1}$$

for which it is a Banach space. That is, the space is a vector space; the norm satisfies the triangle inequality; $\|f\| = 0$ is possible only if $f = 0$; finally, all Cauchy sequences converge: if we have a sequence with entries $f_n \in \mathcal{C}^m(I)$ for which $\|f_n - f_{n'}\|$ can be made arbitrarily small simply by choosing n, n' sufficiently large, then the f_n (and all their derivatives up to the mth) converge uniformly to some function f in $\mathcal{C}^m$ (and its derivatives).

In the case of a multivariate domain $\Omega \subset \mathbb{R}^d$, we define $\mathcal{C}^m(\Omega)$ to be the space of continuous functions which have bounded and continuous partial derivatives $\partial^\alpha f$, $|\alpha| := |\alpha_1| + \cdots + |\alpha_d| \le m$. This space can also be equipped with the norm

$$\|f\|_{\mathcal{C}^m(\Omega)} := \sup_{x\in\Omega}|f(x)| + \sum_{|\alpha|=m}\sup_{x\in\Omega}|\partial^\alpha f(x)|. \tag{A2}$$

for which it is a Banach space.

In many instances, one is somehow interested in measuring smoothness in an average sense: for this purpose it is natural to introduce the *Sobolev spaces* $W^{m,p}(\Omega)$ consisting of all functions $f \in \mathcal{L}^p$ with partial derivatives up to order m in $\mathcal{L}^p$. Here p is a fixed index in $[1, +\infty]$. (Recall that $\|f\|_{\mathcal{L}^p} = [\int_\Omega |f(x)|^p]^{1/p}$ if $p < +\infty$ and $\|f\|_{\mathcal{L}^\infty} = \sup_{x\in\Omega}|f(x)|$.) This space is also a Banach space, when equipped with the norm

$$\|f\|_{W^{m,p}} := \|f\|_{\mathcal{L}^p} + \sum_{|\alpha|=m}\|\partial^\alpha f\|_{\mathcal{L}^p}. \tag{A3}$$

Note that the norm (A2) for $\mathcal{C}^m$ spaces coincides with the $W^{m,\infty}$ norm.

All the above spaces share the common feature that the regularity index is an integer. In many applications, one is interested in allowing fractional orders of smoothness, in order to describe the regularity of a function in a more precise way. The question thus arises of *how to fill the gaps between integer smoothness classes.* There are at least two instances where such a generalization is very natural.

- In the case of $\mathcal{L}^2$-Sobolev spaces $H^m := W^{m,2}$ and when $\Omega = \mathbb{R}^d$, we can define an equivalent norm based on the Fourier transform, since by Parseval's formula we have

$$\|f\|^2_{H^m} \sim \int_{\mathbb{R}^d}(1 + |\omega|^{2m})|\hat{f}(\omega|^{2m})\,d\omega. \tag{A4}$$

  For a noninteger $s \ge 0$, it is thus natural to define the space H^s as the set of all $\mathcal{L}^2$ functions such that

$$\|f\|^2_{H^s} := \int_{\mathbb{R}^d}(1 + |\omega|^{2s})|\hat{f}(\omega)|^2 d\omega \tag{A5}$$

  is finite.

- In the case of $\mathcal{C}^m$ spaces, we note that

$$\sup_{x\in\Omega}|f(x) - f(x-h)| \le C|h| \tag{A6}$$

  if $f \in \mathcal{C}^1$ for any $h \in \mathbb{R}^d$, whereas for an arbitrary function $f \in \mathcal{C}^0$

$$\sup_{x\in\Omega}|f(x) - f(x-h)|$$

  might go to zero arbitrarily slowly as $|h| \to 0$. This motivates the definition of the Hölder space $\mathcal{C}^s$, $0 < s < 1$ consisting of those $f \in \mathcal{C}^0$ such that

$$\sup_{x\in\Omega}|f(x) - f(x-h)| \le C|h|^s. \tag{A7}$$

  If $m < s < m+1$, a natural definition of $\mathcal{C}^s$ is given by $f \in \mathcal{C}^m$ and $\partial^\alpha f \in \mathcal{C}^{s-m}$, $|\alpha| = m$. It is not difficult to prove that this property can also be expressed as

$$\sup_{x\in\Omega}|\Delta^n_h f(x)| \le C|h|^s \tag{A8}$$

  where $n > s$ and Δ^n_h is the nth-order finite-difference operator defined recursively by $\Delta^1_h f(x) = f(x) - f(x-h)$ and $\Delta^n_h f(x) = \Delta^1_h(\Delta^{n-1}_h)f(x)$ (for example $\Delta^2_h f(x) = f(x) - 2f(x-h) + f(x-2h)$). When s is not an integer, the spaces $\mathcal{C}^s$ that we have defined are also denoted as $W^{s,\infty}$.

The definition of "order of smoothness s in $\mathcal{L}^p$" for s noninteger and p different from 2 or ∞ is more subject to arbitrary choices. Among others, one can find the following.

- Sobolev spaces $W^{s,p}$ defined (if $m < s < m+1$) by

$$\|f\|_{W^{s,p}} = \|f\|_{\mathcal{L}^p} + \sum_{|\alpha|=m}\left[\int_{\Omega\times\Omega}\frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|^p}{|x-y|^{(s-m)p+d}}\,dx\,dy\right]^{1/p}. \tag{A9}$$

  These spaces coincide with those defined by means of Fourier transform when $p = 2$ (see [1] for a general treatment).

- Bessel-potential spaces $H^{s,p}$ defined by means of the Fourier transform operator $\mathcal{F}$

$$\|f\|_{H^{s,p}} = \|f\|_{\mathcal{L}^p} + \|\mathcal{F}^{-1}(1 + |\cdot|^s)\mathcal{F}f\|_{\mathcal{L}^p}. \tag{A10}$$

  These spaces coincide with the Sobolev spaces $W^{m,p}$ when m is an integer and $1 < p < +\infty$ (see [23, p. 38]), but their definition by (A10) requires that $\Omega = \mathbb{R}^d$ in order to apply the Fourier transform.

- Besov spaces $B^s_{p,q}$, involving an extra parameter q that we define later through finite differences. These spaces include most of those that we have listed so far as particular cases. As we shall see, one of their main features is that they can be exactly characterized by multiresolution approximation error, as well as by the size properties of the wavelet coefficients.

We define the nth-order $\mathcal{L}^p$ modulus of smoothness of f by

$$\omega_n(f, t)_{\mathcal{L}^p} = \sup_{|h|\le t}\|\Delta^n_h f\|_{\mathcal{L}^p(\Omega_{h,n})} \tag{A11}$$

where

$$\Omega_{h,n} := \{x \in \Omega\,;\, x - kh \in \Omega,\, k = 0, \ldots, n\};$$

(A11) thus expresses that we measure the "size" of $\Delta_h^n f$ in the $\mathcal{L}^p$-norm, where we restrict to $\mathcal{L}^p(\Omega_{h,n})$ to ensure that all the arguments $x-kh$ occurring in the computation of $\Delta_h^n f(x)$ still live in Ω. For $p,\ q \geq 1$, $s > 0$, the Besov space $B^s_{p,q}$ consists of those functions $f \in \mathcal{L}^p$ such that

$$(2^{sj}\omega_n(f, 2^{-j})_{\mathcal{L}^p})_{j\geq 0} \in \ell^q. \tag{A12}$$

Here, n is an integer strictly larger than s. A natural norm for such a space is given by

$$\|f\|_{B^s_{p,q}} := \|f\|_{\mathcal{L}^p} + \|(2^{sj}\omega_n(f, 2^{-j})_{\mathcal{L}^p})_{j\geq 0}\|_{\ell^q}. \tag{A13}$$

Note that this expression *a priori* depends on the choice of n. However, it can be shown that for two values $n_1,\ n_2 > s$ one obtains equivalent norms. If $q = \infty$, the condition (A12) simply means that

$$\|\Delta_h^n f\|_{\mathcal{L}^p} \leq Ch^{-s}, \qquad \text{for } |h| \leq 1.$$

For $q < \infty$, the decay condition on $\Delta_h^n f$ is slightly stronger, since we require that the sequence $(2^{sj}\omega_n(f, 2^{-j})_{\mathcal{L}^p})_{j\geq 0}^q$ be summable. The space $B^s_{p,q}$ also represents "s order of smoothness measured in $\mathcal{L}^p$"; the parameter q allows a finer tuning on the degree of smoothness—one has $B^s_{p,q_1} \subset B^s_{p,q_2}$ if $q_1 \leq q_2$—but plays a minor role in comparison to s since clearly

$$B^{s_1}_{p,q_1} \subset B^{s_2}_{p,q_2}, \qquad \text{if } s_1 \geq s_2 \tag{A14}$$

regardless of the values of q_1 and q_2. Roughly speaking, smoothness of order s in $\mathcal{L}^p$ is expressed here by the fact that, for n large enough, $\omega_n(f, t)_{\mathcal{L}^p}$ goes to 0 like $\mathcal{O}(t^s)$ as $t \to 0$.

Clearly, $\mathcal{C}^s = B^s_{\infty,\infty}$ when s is not an integer. It can also be proved that when s is not an integer, $W^{s,p} = B^s_{p,p}$. These spaces are different from one another for integer values of s, except when $p = 2$ in which case $H^s = B^s_{2,2}$ for all values of s (see [23, p. 38]).

Sobolev, Besov, and Bessel-potential spaces satisfy two simple embedding relations.

- For fixed p (and arbitrary q in the case of Besov spaces, see (A14)), the spaces get larger as s decreases.
- In the case where Ω is a bounded domain, for fixed s (and fixed q in the case of Besov spaces), the spaces get larger as p decreases, since $\|f\|_{\mathcal{L}^{p_1}} \leq C\|f\|_{\mathcal{L}^{p_2}}$ if $p_1 \leq p_2$.

A less trivial type of embedding is known as *Sobolev embedding.* In the case of Sobolev spaces, it states that

$$W^{s_1,p_1} \subset W^{s_2,p_2} \qquad \text{if } s_1 - s_2 \geq d(1/p_1 - 1/p_2) \tag{A15}$$

except in the case where $p_2 = +\infty$ and $s_1 - d(1/p_1 - 1/p_2)$ is an integer, for which one needs to assume $s_1 - s_2 > d(1/p_1 - 1/p_2)$. For example, in the univariate case, any H^1 function has also $\mathcal{C}^{1/2}$ smoothness. In the case of Besov spaces, the embedding relation is given by

$$B^{s_1}_{p_1,p_1} \subset B^{s_2}_{p_2,p_2} \qquad \text{if } s_1 - s_2 \geq d(1/p_1 - 1/p_2) \tag{A16}$$

with no restrictions on the indexes $s_1,\ s_2 \geq 0$ and $p_1,\ p_2 \geq 1$. The proof of these embeddings can be found in [1] for Sobolev spaces and in [23] for Besov spaces.

As an exercise, let us see how these embeddings can be used to derive the range of r such that $B^r_{2,q}([0, 1])$ can contain discontinuous functions. If $r > 1/2$, then there exists $\varepsilon > 0$ such that $r - 2\varepsilon > 1/2$; applying first (A14) and then (A16) we find that $B^r_{2,q} \subset B^{r-\varepsilon}_{2,2} \subset B^{\varepsilon}_{\infty,\infty} = \mathcal{C}^{\varepsilon}$, so all functions in $B^r_{2,q}$ are continuous. Therefore, only $B^r_{2,q}$ with $r \leq 1/2$ can contain discontinuous functions.

As mentioned earlier, Besov spaces can also be characterized by approximation properties. More precisely, if f_j denotes the projection of f onto the space $\mathcal{V}_j$ ($f_{-1} = 0$), under certain assumptions that we shall discuss later, the Besov norm $\|f\|_{B^s_{p,q}}$ is equivalent to

$$\|(2^{sj}\|f - f_j\|_{\mathcal{L}^p})_{j\geq 0}\|_{\ell^q} \tag{A17}$$

or to

$$\|(2^{sj}\|f_j - f_{j-1}\|_{\mathcal{L}^p})_{j\geq 0}\|_{\ell^q}. \tag{A18}$$

If we have at our disposal a d-dimensional wavelet basis to characterize the details $f_j - f_{j-1} = \sum_{k\in K_j} c_{j,k}\psi_{j,k}$, then we can use the equivalence

$$\|f_j - f_{j-1}\|_{\mathcal{L}^p} \sim 2^{(d/2-d/p)j}\|(c_{j,k})_{k\in K_j}\|_{\ell^p}$$

at each level to prove a third equivalent norm in terms of the wavelet coefficients

$$\|(2^{sj}2^{(d/2-d/p)j}\|(c_{j,k})_{k\in K_j}\|_{\ell^p})_{j\geq 0}\|_{\ell^q} \tag{A19}$$

(for notational simplicity, we incorporate in (A19) the coarse scale approximation coefficients of f_0 in the set $(c_{0,k})_{k\in K_0}$).

These equivalences mean that the modulus of smoothness $\omega_n(f, 2^{-j})_{\mathcal{L}^p}$ in the definition of $B^s_{p,q}$ can be replaced either by $\|f - f_j\|_{\mathcal{L}^p}$ or by $\|f_{j+1} - f_j\|$. The equivalence of (A17) or (A18) to the Besov norm $\|f\|_{B^s_{p,q}}$ follows whenever the spaces $\mathcal{V}_j$ satisfy the following two assumptions.

- The $\mathcal{V}_j$ must satisfy an approximation property that takes the form of a *direct estimate*

  $$\|f - f_j\|_{\mathcal{L}^p} \leq C\omega_n(f, 2^{-j})_{\mathcal{L}^p}. \tag{A20}$$

  Such an estimate ensures that a smooth function has a fast rate of approximation.
- They must also satisfy smoothness properties that take the form of an *inverse estimate*

  $$\omega_n(f, t)_{\mathcal{L}^p} \leq C[\min(1, t2^j)]^n\|f\|_{\mathcal{L}^p}, \qquad \text{if } f \in \mathcal{V}_j. \tag{A21}$$

  Such an estimate takes into account the smoothness of the spaces $\mathcal{V}_j$: it ensures that a function that is approximated at a sufficiently fast rate by these spaces should also have some smoothness.

One can show that the direct estimate is satisfied if and only if all polynomials up to order $n-1$ can be written as combinations of functions in $\mathcal{V}_j$, or equivalently (in the wavelet case) if the wavelets have n vanishing moments. On the other hand, the inverse estimate requires that the scaling function φ that generates $\mathcal{V}_j$ is smooth, more precisely that φ (and, therefore, also all the wavelets) is in $W^{n,p}$. Note that the direct estimate immediately implies that (A17) is less than $\|f\|_{B^s_{p,q}}$. A more refined mechanism, using the inverse estimate (as well as some discrete Hardy inequalities) is used to prove the full equivalence between

$||f||_{B^s_{p,q}}$, (A17), and (A18). We refer to [5, Ch. III] for a detailed proof of these results.

These equivalences show that the distortion rate $N^{-t/d}$ ($N = \dim(\mathcal{V}_j)$) can be achieved by a linear multiscale approximation process, if and only if the function has roughly "t derivatives in $\mathcal{L}^p$," a generalization of the result for $p = 2$ that we used in Section II-A.

An instance of a nonlinear approximation result was addressed in Section II-A (see (13) and Proposition 3), in the case where the error is measured in $\mathcal{L}^2$. Here also a more general result, proved in [9], holds in the case where the error is measured in $\mathcal{L}^p$, $1 < p < +\infty$. In that case, it is natural to define

$$\mathcal{A}_N f = \sum_{(j,k)\in E^p_N(f)} c_{j,k}\psi_{j,k} \tag{A22}$$

where $E^p_N(f)$ is the set of indexes corresponding to the N largest contributions in the $\mathcal{L}^p$ metric of the wavelet expansion of f, i.e., $\#(E^p_N(f)) = N$ and $||c_{j,k}\psi_{j,k}||_{\mathcal{L}^p} \geq ||c_{l,m}\psi_{l,m}||_{\mathcal{L}^p}$ if $(j, k) \in E^p_N(f)$ and $(l, m) \notin E^p_N(f)$. The result of DeVore, Jawerth, and Popov is that $||f - \mathcal{A}_N f||_{\mathcal{L}^p} \sim N^{-s/d}$ is achieved for functions $f \in B^s_{q,q}$ where $1/q = 1/p + s/d$.

Note that this relation between p and q corresponds to a critical case of the Sobolev embedding of $B^s_{q,q}$ into $\mathcal{L}^p$. In particular, $B^s_{q,q}$ is not contained in $B^\varepsilon_{p,p}$ for any $\varepsilon > 0$, so that *no decay rate can be achieved by a linear approximation process* for all the functions f in the space $B^s_{q,q}$. (For *some* functions in $B^s_{q,q}$, which happen to also lie in spaces for which an independent linear approximation theorem can be written, it is of course possible to get a linear approximation rate; the point here is that this is possible only via such additional information.)

Note also that for large values of s, the parameter q given by $1/q = 1/p + s/d$ is smaller than 1. In such a situation, the space $B^s_{q,q}$ is not a Banach space any more and (A13) is only a quasi-norm (it fails to satisfy the triangle inequality $||x + y|| \leq ||x|| + ||y||$). However, this space is still contained in $\mathcal{L}^1$ (by a Sobolev-type embedding) and its characterization by means of wavelets coefficients according to (A19) still holds. Letting q go to zero as s goes to infinity allows the presence of singularities in the functions of $B^s_{q,q}$ even when s is large: for example, a function which is piecewise $\mathcal{C}^n$ on an interval except at a finite number of isolated points of discontinuities belongs to all $B^s_{q,q}$ for $q < 1/s$ and $s < n$. This is a particular instance where a nonlinear approximation process will perform substantially better than a linear projection.

ACKNOWLEDGMENT

The authors wish to thank Ron DeVore, David Donoho, Brad Lucier, and Stéphane Mallat for fruitful discussions.

REFERENCES

[1] R. Adams, *Sobolev Spaces*. New York: Academic, 1975.
[2] J. P. D'Ales and A. Cohen, "Non-linear approximation of random functions," *SIAM J. Appl. Math*, vol. 57, no. 2, pp. 518–540, 1997.
[3] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based signal processing using hidden Markov models," *IEEE Trans. Signal Processing (Special Issue on Theory and Application of Filter Banks and Wavelet Transforms)*, vol. 46, pp. 886–902, Apr. 1998.
[4] T. Berger, *Rate Distortion Theory*, NJ, Englewood Cliffs: Prentice-Hall, 1971.
[5] A. Cohen, "Wavelets and multiscale decompositions in numerical analysis," in *Handbook of Numerical Analysis*, P. G. Ciarlet and J. L. Lions, Eds. Amsterdam, The Netherlands: Elsevier, 1999.
[6] A. Cohen, W. Dahmen, I. Daubechies, and R. DeVore, "Kolmogorov entropies and wavelet-based coding," Dept. Appl. Math., Princeton Univ., Princeton, NJ, 1998.
[7] I. Daubechies, *Ten Lectures on Wavelets*, ser. CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia, PA: SIAM, 1988, vol. 61.
[8] R. DeVore, "Nonlinear approximation," *Acta Numer.*, 1998, to be published.
[9] R. DeVore, B. Jawerth, and V. Popov, "Compression of wavelet decompositions," *Amer. J. Math.*, vol. 114, pp. 737–785, 1992.
[10] R. DeVore and G. Lorentz, *Constructive Approximation*. Berlin, Germany: Springer-Verlag, 1993.
[11] R. DeVore and X. M. Yu, "Degree of adaptive approximation," *Math. Comp.*, vol. 55, no. 192, pp. 625–635, 1990.
[12] D. Donoho, "Unconditional bases and bit level compression," *Appl. Comp. Harm. Anal.*, vol. 3, pp. 388–392, 1996.
[13] F. Falzon and S. Mallat, "Analysis of low bit rate image transform coding," *IEEE Trans. Signal Processing (Special Issue on Theory and Application of Filter Banks and Wavelet Transforms)*, vol. 46, pp. 1027–1042, Apr. 1998.
[14] R. H. Jonsson and R. M. Mersereau, "Subband coding of video using adaptive quantization," in *Proc. Int. Symp. Circuits and Systems*, 1994, pp. 3.285–3.288.
[15] R. Joshi, H. Jafarkhani, J. Kasner, T. Fischer, N. Farvardin, M. Marcellin, and R. Bamberger, "Comparison of different methods of classification in subband coding of images," *IEEE Trans. Image Processing*, vol. 6, pp. 1473–1487, Nov. 1997.
[16] ITU/ISO, "Information Technology—Digital Compression and Coding of Continuous-Tone Still Images," std., JPEG ITU-T Rec. T.81-ISO/IEC no. 10918-1, 1993.
[17] L. J. Kleinwaks, "Nonlinear wavelet approximation of multidimensional piecewise smooth stochastic processes," Math. Dept., Princeton Univ., Princeton, NJ, Project Rep. for MAT584, Spring 1998.
[18] S. M. LoPresto, K. Ramchandran, and M. T. Orchard, "Image coding based on mixture modeling of wavelet coefficients and a fast estimation-quantization framework," in *Proc. IEEE Data Compression Conf. (DCC'97)*, Mar. 1997, pp. 221–230.
[19] Y. Meyer, *Ondelettes et Opérateurs*. Paris, France: Hermann, 1990.
[20] ISO, "Information Technology-Generic Coding of Moving Pictures and Associated Audio Information: Video," std., MPEG: ISO/IEC 13818-2:1996(E) 1996-05-15.
[21] A. Said and W. A. Pearlman, "A new fast and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, pp. 243–250, June 1996.
[22] J. Shapiro, "Embedded image coding using zero-trees of wavelet coefficients," *IEEE Trans. Signal Processing*, vol. 41, pp. 3445–3462, 1993.
[23] H. Triebel, *Theory of Function Spaces*. Basel, Switzerland: Birkhauser, 1983.
[24] M. J. Tsai, J. Villasenor, and F. Chen, "Stack-run image coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, pp. 519–521, Oct. 1996.
[25] J. Villasenor. Wavelet image coding: PSNR results. [Online]. Available: http://www.icsl.ucla.edu~ipl//psnr_results.html
[26] Z. Xiong, O. Guleryuz, and M. T. Orchard, "A DCT-based embedded image coder," *IEEE Signal Processing Lett.*, vol. 3, pp. 289–290, Nov. 1996.
[27] Z. Xiong, K. Ramchandran, and M. T. Orchard, "Space-frequency quantization for wavelet image coding," *IEEE Trans. Image Processing*, vol. 6, pp. 677–693, May 1997.