# Optimal Tree Approximation with Wavelets

*Richard Baraniuk*

Department of Electrical and Computer Engineering
Rice University
Houston, Texas 77005, USA

## ABSTRACT

The more a priori knowledge we encode into a signal processing algorithm, the better performance we can expect. In this paper, we overview several approaches to capturing the structure of singularities (edges, ridges, etc.) in wavelet-based signal processing schemes. Leveraging results from approximation theory, we discuss nonlinear approximations on trees and point out that an optimal tree approximant exists and is easily computed. The optimal tree approximation inspires a new hierarchical interpretation of the wavelet decomposition and a tree-based wavelet denoising algorithm that suppresses spurious noise bumps.

**Keywords:** Wavelets, trees, nonlinear approximation, Besov space,optimization

## 1. INTRODUCTION

The wavelet transform provides a natural setting for developing new signal and image processing algorithms, especially for signals and images rich in *singularities* (edges, ridges, and other transients). Since wavelets form a basis,[1,2] they can reproduce arbitrary functions, from highly structured real-world signals and images to completely unstructured noise. In linguistic terms, the wavelet *vocabulary* can be too expressive.[3] * To better process real-world signals and images, we must narrow this vocabulary's scope by imposing a set of constraints — a *grammar* — that captures the salient structures of singularities. While much research has concentrated on developing new wavelet vocabularies, until recently relatively little effort has gone into grammatical modeling.

Clearly, the more knowledge we have about the structure of a signal class, the more accurate the model we can construct and the better the performance of any signal processing algorithm derived from it. But what structure to exploit in wavelet transforms?

The wavelet transform of a 1-d signal consists of the wavelet coefficients $\{w_{j,k}\}$, indexed by a scale parameter $j$ and a space parameter $k$ (higher dimensions are handled similarly).[1,2] The wavelet coefficients form a pyramid or tree that represents signal structure from coarse to fine scales. Singularities are particularly simply represented: the energy from a singularity localizes along one branch of the tree (see Figure 1).[1] To best process singularity-rich data, it is key that we match this persistence of singularity energy in the wavelet tree. Very roughly, we can delineate between *statistical* and *deterministic* tree-based modeling approaches.[1,7–15]

In this paper, we will focus on an optimization-based approach to tree modeling in the wavelet domain that extends the concept of *nonlinear approximation*[16,17] to a kind of *optimal tree approximation.* Our results extend those of Cohen et al[18,19] somewhat but are closely related to those of Donoho et al[20,21] and Engel.[22] In a sense, this paper is about different orderings of the terms in the wavelet series.

After quickly reviewing wavelets in Section 2 and linear and nonlinear approximation in Section 3, we turn to tree approximation in Section 4. In Section 5 we demonstrate how our ideas (and those of Donoho et al[20,21] and Engel[22]) can improve upon wavelet denoising by taking into account the persistency properties of singularities. We close in Section 6 with a discussion and conclusions.

This work was commenced at the Isaac Newton Institute of Cambridge University and was supported by the Rosenbaum Fellowship and by NSF grant MIP–9457438, DARPA/AFOSR grant F49620–97–1–0513, ONR grant N00014–99–1–0813, and the Texas Instruments Leadership University Program.

Email: richb@rice.edu; Internet: www.dsp.rice.edu

[*] While we will use these linguistic terms loosely, the analogy with signal modeling can be made formal; see refs[4–6] for examples.

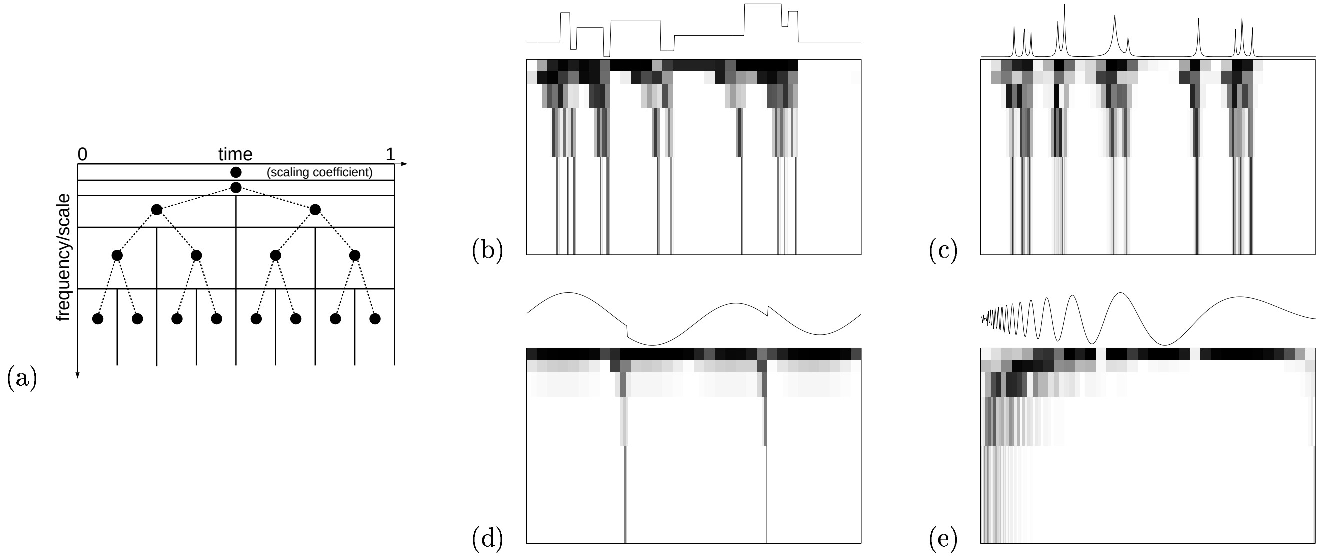


**Figure 1.** *(a) The coefficients of the wavelet transform naturally form a binary tree structure flowing from the coarsest scale (root) to the finest scale (leaves). Each black node represents a wavelet coefficient. Image plots of the wavelet coefficients $\{w_{j,k}\}$ for Donoho's four test signals (b) Blocks, (c) Bumps, (d) HeaviSine, and (e) Doppler demonstrate the strong* **persistence across scale** *property of large and small wavelet coefficient values. Each rectangular tile corresponds to one coefficient; its darkness corresponds to the size of $|w_{j,k}|$, with white meaning $|w_{j,k}| = 0$.*

## 2. WAVELET TRANSFORM

Given a lowpass scaling function $\phi$ and bandpass wavelet function $\psi$, define the multiscale atoms

$$\phi_{j,k}(t) := 2^{j/2}\,\phi(2^j t - k), \qquad \psi_{j,k}(t) := 2^{j/2}\,\psi(2^j t - k), \tag{1}$$

with $j$ and $k$ indexing their scale and position, respectively. For special choices of $\phi$ and $\psi$, these atoms form an orthonormal basis for $L_2$,[1,2] and we have the following multiscale representation of a signal $f$

$$f \;=\; \sum_k u_{j_0}\phi_{j_0,k} \;+\; \sum_{j \geq j_0}\sum_k w_{j,k}\psi_{j,k}, \tag{2}$$

with $u_{j_0,k} := \langle f, \phi_{j_0,k}\rangle$ and $w_{j,k} := \langle f, \psi_{j,k}\rangle$. We will emphasize signals in $L_2([0,1])$ but will not delve into the details of the required boundary adapted wavelets.[1,2] In this case, $j_0 \geq 0$, and there are $2^j$ wavelet atoms per scale $j$. When convenient, we will use the short-hand notation $I := (j,k)$ as a multi-index for the atoms; $I$ corresponds to the interval $2^{-j}[k, k+1)$, the support interval of the Haar wavelet $\psi_I$. Images and higher-dimensional data can be handled similarly using tensor products of $\phi$ and $\psi$.

The multiscale nesting structure of the wavelet atoms — the support of each $\psi_{j,k}$ contains the supports of $\psi_{j+1,2k}$ and $\psi_{j+1,2k+1}$ — induces a binary *tree structure* on the wavelet coefficients (see Figure 1(a)). We say that $w_{j,k}$ is the *parent* of its two *children* $w_{j+1,2k}$ and $w_{j+1,2k+1}$. For simplicity (but without loss of generality, see Section 4.4), we will assume in most of the sequel that $j_0 = 0$, which leads to a single wavelet tree with *root* at the coarsest-scale wavelet coefficient $w_{0,0}$ (as in Figure 1(a)). The single scaling coefficient $u_{0,0}$ crowns the tree. Define a *subtree* as any connected set of nodes in the tree; define a *rooted subtree* as a subtree that includes the main tree root.

Wavelets act as multiscale edge detectors: Large/small wavelet coefficients indicate the presence of an edge/smooth region in the support of the wavelet. Due to the nesting property of child wavelets inside their parents, edges and other singularities manifest themselves in the wavelet domain as chains of large wavelet coefficients along the branches of the tree (recall Figure 1).

Many of the remarkable properties of wavelets stem from the fact that they form unconditional bases for a plethora of function spaces, including the $L_p$, $1 < p < \infty$, Sobolev, Besov, and Triebel spaces.[16,17,19,23] Unconditionality gives wavelet bases optimal energy compaction properties: For signals in these spaces, the wavelet representation packs more energy into a fixed number of coefficients than any other basis.

Roughly speaking, the *Besov space* $B^s_q(L_p)$ contains functions with $s > 0$ derivatives as measured in $L_p$, $p > 0$ (the $q$ parameter makes finer distinctions in smoothness). For wavelets with $r > s$ vanishing moments, the Besov norm is equivalent to a simple sequence norm on the wavelet coefficients[16,17,19,23]:

$$\|f\|_{B^s_q(L_p)} \asymp \|u_{j_0,k}\|_p + \left(\sum_{j\geq j_0}\left(2^{js'}\left(\sum_k |w_{j,k}|^p\right)^{1/p}\right)^q\right)^{1/q}, \tag{3}$$

with $s' = s - 1/p + 1/2$ and the obvious modifications if $p$ or $q = \infty$. This norm comprises an $l_p$ norm within each scale $j$ and then a weighted $l_q$ norm across scale. Clearly a signal lies in a given Besov space if and only if its wavelet coefficients decay sufficiently rapidly (with the decay constraints becoming more stringent as $s$ and $p$ increase). Special cases of Besov spaces include the homogeneous Besov spaces $B^s_p(L_p)$ and the Sobolev spaces $W^s(L_2) := B^s_2(L_2)$. When $s = 1/p - 1/2$, the $B^s_p(L_p)$ norm has a particularly simple form:

$$\|f\|_{B_p^{1/p-1/2}(L_p)} \asymp \|u_{j_0,k}\|_p + \|w_{j,k}\|_p. \tag{4}$$

## 3. LINEAR VS. NONLINEAR APPROXIMATION

Wavelet approximation deals with the following problem: approximate a function $f$ using $n$ terms of its wavelet series representation (2). (We will employ all of the scaling coefficient terms $\sum_k u_{j_0,k}\phi_{j_0,k}$ in our approximations without including them in the total $n$.)

### 3.1. Linear approximation

In *linear approximation*, we construct the approximant $\widehat{f}^n_{\mathsf{L}}$ by projecting $f$ onto a fixed $n$-dimensional subspace of $L_2$ defined by $n$ fixed wavelets. Typically, we employ the first $n$ terms in the double sum of (2):

$$\widehat{f}^n_{\mathsf{L}} := \sum_k u_{j_0}\phi_{j_0,k} + \underbrace{\sum_{j_0\leq j\leq J}\sum_k w_{j,k}\psi_{j,k}}_{n \text{ terms}}. \tag{5}$$

That is, we use wavelets up to some fixed scale $J < \infty$. Each signal is approximated by the *same* $n$ wavelet terms, which makes the process linear: $(\widehat{f_1 + f_2})^n_{\mathsf{L}} = (\widehat{f}_1)^n_{\mathsf{L}} + (\widehat{f}_2)^n_{\mathsf{L}}$. Linear approximation is hierarchical in the sense that as $n \to \infty$, we add resolution to the approximation one scale band at a time.

For any function $f \in W^s(L_2)$, the $L_2$ error of approximation can be bounded as[16,17,19†]

$$\|f - \widehat{f}^n_{\mathsf{L}}\|_2 \leq C_{\mathsf{L}} n^{-s}. \tag{6}$$

### 3.2. Nonlinear approximation

In *nonlinear approximation* — also known as *n-term* or *greedy* approximation — we order the terms in the approximant not according to wavelet scale $j$ but according to wavelet coefficient *size* $|w_I|^2$. That is, for $j \geq j_0$ we order the wavelet coefficients such that

$$|w_{I_1}|^2 \geq |w_{I_2}|^2 \geq \ldots, |w_{I_i}|^2 \geq |w_{I_{i+1}}|^2 \geq \cdots \tag{7}$$

and form the approximant using the first $n$ (largest) elements in this list:

$$\widehat{f}^n_{\mathsf{N}} := \sum_k u_{j_0}\phi_{j_0,k} + \sum_{i=1}^n w_{I_i}\psi_{I_i}. \tag{8}$$

In general, different signals will be represented using different wavelet terms, which makes the process nonlinear: $(\widehat{f_1 + f_2})^n_{\mathsf{N}} \neq (\widehat{f}_1)^n_{\mathsf{N}} + (\widehat{f}_2)^n_{\mathsf{N}}$. Nonlinear approximation is hierarchical in the sense that wavelet terms are included in order of size, with large coefficients first and small coefficients later.

[†] Similar results hold when we measure the error in other $L_p$ norms.[16]

Since we use the largest $n$ terms in the representation (2), $\|f - \widehat{f}^n_{\mathsf{N}}\|_2 \leq \|f - \widehat{f}^n_{\mathsf{L}}\|_2$. Furthermore, for any function $f \in B^s_p(L_p)$, $s \geq 1/p - 1/2$, the nonlinear approximation error can be bounded as[16,17,19]

$$\|f - \widehat{f}^n_{\mathsf{N}}\|_2 \leq C_{\mathsf{N}} n^{-s}, \tag{9}$$

with $C_{\mathsf{N}} \leq C_{\mathsf{L}}$. This might not seem much of an improvement over linear approximation (see (6)); however this is not the case, for two reasons. First, for a given $s$, the space $B^s_p(L_p)$ is considerably larger than $W^s(L_2)$, meaning that the nonlinear scheme can approximate more functions at rate $n^{-s}$. Second, if $f \in W^s(L_2)$ then also $f \in B^a_b(L_b)$, $a = 1/b - 1/2$, with $a \geq s$, meaning that $f$ has greater smoothness in $B^a_b(L_b)$ (and hence faster error decay with nonlinear approximation) than in $W^s(L_2)$ (with linear approximation).[17]

**Relation to thresholding:** Given $N$ total wavelet terms in (2) (as we would have in the representation of a discrete-time signal), assembling the $n$-term approximation requires that we sort the wavelet coefficients. The cost of sorting — $O(N \log N)$ in general — exceeds the $O(N)$ cost of the forward and inverse wavelet transforms.

Applying a (hard) threshold $\tau$ to the wavelet coefficients[‡]

$$\widehat{w}_I = \begin{cases} w_I, & |w_I|^2 > \tau \quad \text{(keep)} \\ 0, & |w_I|^2 \leq \tau \quad \text{(kill)} \end{cases} \tag{10}$$

also generates the coefficients a nonlinear approximation. And, in contrast to sorting, thresholding is an $O(N)$ operation.

Note, however, that the exact relationship between the threshold $\tau$ and the number of terms $n$ in the approximation is signal dependent, and hitting a prespecified $n$ terms using thresholding will require multiple thresholding passes in general. If more than $\log N$ thresholdings are required, then sorting is a more efficient option. Nevertheless, for signal estimation (denoising)[23] and certain compression applications,[18,19,24,25] it is more natural to use $\tau$ as a control parameter than $n$.

## 4. TREE APPROXIMATION

Nonlinear approximation, while optimal in a squared error sense, does not explicitly exploit the structure of the $n$ largest wavelet terms it employs. There are several justifications for taking this structure into account. In coding applications, for example, we must encode not only the $n$ coefficients we wish to transmit, but also their locations in the wavelet tree. Clearly if the $n$ coefficients lie on an unstructured set in the wavelet domain, then the cost of position encoding could exceed value encoding.[18,19] We will see another example involving wavelet denoising below in Section 5.[20–22]

In *tree approximation*, we seek a nonlinear approximation possessing both a small error and substantial structure.[18,19] But what structure to exploit? Large signal classes are known to exhibit considerable structure in the wavelet domain. Wavelet coefficients of smooth signals (those in Besov spaces, for example) must decay as scale $j \to \infty$. Wavelet coefficients of singularities have large wavelet coefficients that persist along the branches of the wavelet tree.[1,11,24] Both of these types of signal behaviors (smooth regions and singularities) lead to wavelet coefficients that are large on rooted subtrees of the main wavelet tree. The configuration of such a tree is easily encoded (most simply by the locations of its leaves, for example).[18,19,24,25]

A tree approximation represents a function by the largest wavelet coefficients in a connected, rooted subtree of the main wavelet tree. That is, a wavelet coefficient is not included in the approximant before any of its ancestors. Depending on how we define "largest" we obtain different approximations. Surprisingly, the performance of tree approximation comes very close to that of $n$-term nonlinear approximation for a large class of signals.

### 4.1. Greedy tree approximation

To be clear in the sequel, we will add the qualifier "scalar" when referring to $n$-term nonlinear approximation à la Section 3.2. Define $\mathcal{T}$ as a connected rooted subtree of wavelet coefficients containing $|\mathcal{T}|$ wavelet coefficients.

In $n$-term *greedy tree approximation*[§] we simply perform the usual scalar nonlinear approximation of Section 3.2 and then form a rooted tree from the selected (largest) coefficients. More specifically, we (1) find the $m$, $m \leq n$,

[‡] Soft thresholding (where we shrink all $w_I > \tau$ by $\tau$ after thresholding) is advantageous in certain problems.[23]
[§] Called simply "tree approximation" by Cohen et al.[18]

largest wavelet coefficient terms; (2) form the smallest connected, rooted subtree $\mathcal{T}$ that contains all of these $m$ coefficients; and then (3) increase $m$ until $|\mathcal{T}| = n$.[¶] See Figure 2(a)–(c) for a simple example.

Only when the wavelet coefficients decay monotonically along the tree branches toward the leaves will the greedy tree and the scalar nonlinear approximants coincide. In general, the error of the tree approximant will exceed that of the scalar approximant, since the tree approximant can include isolated large coefficients far from the tree root only by including all ancestor coefficients, which may be small (see Figure 2(c)).

Amazingly, for signals in Besov space, tree approximation has essentially the same power as scalar nonlinear approximation. That is, for any function $f \in B_p^s(L_p)$, $s > 1/p - 1/2$ (note the only slightly weakened, strict inequality), the error of greedy tree approximation can be bounded as[18]

$$\|f - \widehat{f}_{\mathsf{G}}^n\|_2 \ \leq C_{\mathsf{G}} n^{-s}, \tag{11}$$

with $C_{\mathsf{G}} \geq C_{\mathsf{N}}$. Tree approximation works well for signals in Besov spaces, because the wavelet coefficients of such signals must decay rapidly with scale, making the likelihood of an isolated large coefficient tree less and less likely as $n \to \infty$.[18,19]

### 4.2. Optimal tree approximation

While greedy tree approximation provides asymptotically near optimal approximation for signals in Besov spaces, it offers no performance guarantees at finite tree sizes $n \ll \infty$. We now discuss an approximation algorithm that provides optimal performance at all tree sizes.

Define the optimal tree (in the $L_2$ error sense) $\mathcal{T}_{\mathsf{O}}$ as the connected rooted tree $\mathcal{T}$ of size $n$ that maximizes the sum of squares of the wavelet coefficients:[‖]

$$\mathcal{T}_{\mathsf{O}} = \arg \max_{|\mathcal{T}|=n} \sum_{I \in \mathcal{T}} |w_I|^2. \tag{12}$$

Define the optimal tree approximant $\widehat{f}_{\mathsf{O}}^n$ as the sum of the wavelet terms with $w_I \in \mathcal{T}_{\mathsf{O}}$. This optimization reduces to a linear programming problem; it can be easily solved in one of two ways.

**Solution 1:** The direct constrained optimization (12) can be solved for a tree of exactly $n$ terms using the "condensing sort and select algorithm" (CSSA).[26,27] Recall that tree approximation coincides with greedy $n$-term approximation (and hence can be solved by simply sorting the wavelet coefficients) when the wavelet data is monotonically nonincreasing along the tree branches out from the root. The CSSA solves (12) in the case of general data by *condensing* the nonmonotonic segments of the tree branches using an iterative sort-and-average routine. In[26,27] the condensed nodes are called "supernodes" (see Figure 2(e)). Condensing a large coefficient far down the tree accounts for the (potentially large) cost of growing the tree to that point.

Since the first step of the CSSA involves sorting all coefficients, overall it requires $O(N \log N)$ computations. However, once the CSSA grows the optimal tree of size $n$, it is extremely cost-effective to grow the optimal trees of sizes $> n$.

While greedy tree approximation provides a tree that is admissible under the constraint $|\mathcal{T}| = n$, it only approximately maximizes the performance measure (12), as Figure 2 shows. It is not enough to choose the largest wavelet coefficient values, because a large value far out in the tree could waste precious tree mass over small coefficients that contribute little to the performance measure.

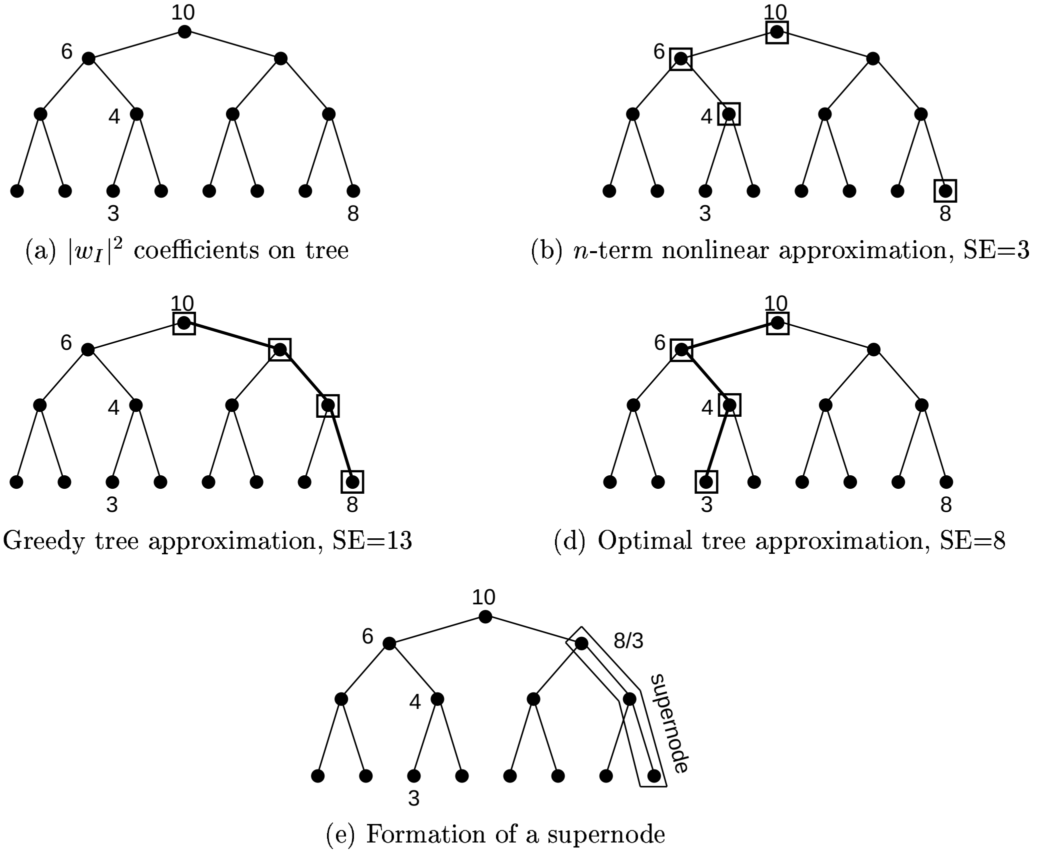
**Solution 2:** The constrained optimization (12) can be rewritten as an unconstrained problem by introducing the Lagrange multiplier $\lambda$[21]

$$\max_{\mathcal{T}} \sum_{I \in \mathcal{T}} |w_I|^2 - \lambda(|\mathcal{T}| - n). \tag{13}$$

Here, $\mathcal{T}$ can be any size. Except for the inconsequential $\lambda n$ term, this optimization coincides with Donoho's "complexity penalized sum of squares",[20,21] which can be solved in only $O(N)$ computations using coarse-to-fine dynamic programming on the tree.

[¶] Note that as we increase $m$ by one, the tree size $n$ will in general grow by some number $\geq 1$.

[‖] This problem was originally posed in the context of designing optimal kernels for time-frequency analysis, but it applies here mutatis mutandis. Donoho et al[20,21] have posed this problem directly as a statistical estimation problem.



(a) $|w_I|^2$ coefficients on tree

(b) $n$-term nonlinear approximation, SE=3

(c) Greedy tree approximation, SE=13

(d) Optimal tree approximation, SE=8

(e) Formation of a supernode

**Figure 2.** *Nonlinear approximation illustrated for $n = 4$ (SE denotes squared error). (a) Wavelet coefficient values $|w_I|^2$ of the signal $f$ placed on the binary wavelet tree. Nodes not labeled carry value zero. (b) The $n$-term nonlinear approximant $\widehat{f}_{\mathsf{N}}^4$ (Section 3.2) employs the largest $n = 4$ wavelet terms, regardless of their position in the tree. (c) The greedy tree approximant $\widehat{f}_{\mathsf{G}}^4$ (Section 3.2) forms the connected, rooted subtree with 4 nodes containing $\widehat{f}_{\mathsf{N}}^m$, for the largest possible $m \leq 4$. In this case, $m = 2$. (d) The optimal tree approximant $\widehat{f}_{\mathsf{O}}^4$ (Section 4.2) forms the connected, rooted subtree with 4 nodes that maximizes the sum of the $|w_I|^2$ in the subtree. (e) It is not optimal to include the $|w_I|^2 = 8$ term in the approximation, because this would waste valuable tree area (two nodes) over zero coefficients. For the purpose of selecting terms in the optimal tree approximation, it is convenient to average the $|w_I|^2 = 8$ term towards the root, forming a "supernode" of value $8/3$.*[26,27]

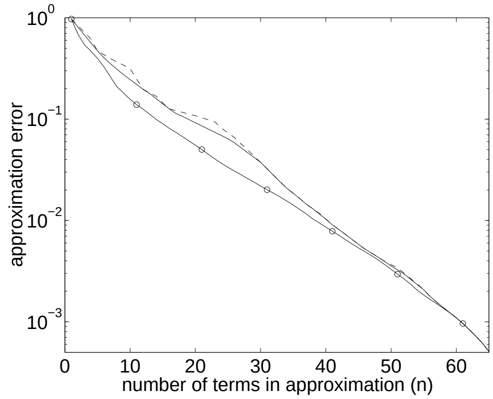


**Figure 3.** *Nonlinear approximation of the doppler test signal by (○) scalar nonlinear, (dash) greedy tree, and (solid) optimal tree approximation. We plot the $L_2$ approximation error on a log scale as a function of the number of terms in the approximation $n$. After $n = 60$, all curves coincide. We transformed the 128-point signal using the Daubechies-4 wavelet*[1,2] *and 7 scales.*

Solutions 1 and 2 play rôles analogous to the "sort" and "threshold" approaches to scalar nonlinear approximation from Section 3.2: The more costly Solution 1 hits the $|\mathcal{T}| = n$ constraint exactly, whereas the more cost-effective Solution 2 exhibits a more complicated, signal-dependent relationship between $\lambda$ and $n$.

Regardless of the solution method, optimal tree approximation outperforms greedy tree approximation in general. Thus, for a function $f \in B_p^s(L_p)$, $s > 1/p - 1/2$, the error of approximation can be bounded as

$$\|f - \widehat{f}_{\mathsf{O}}^n\|_2 \leq C_{\mathsf{O}} n^{-s} \tag{14}$$

with $C_{\mathsf{G}} \geq C_{\mathsf{O}} \geq C_{\mathsf{N}}$. The constant $C_{\mathsf{O}}$ is optimal for this problem.

Figure 3 plots $\|f - \widehat{f}^n\|_2$ for the scalar, greedy tree, and optimal tree approximations to the Doppler test signal. We see from the Figure that the optimal tree error curve forms a lower envelope for the greedy tree error curve. The less monotonic the wavelet coefficients along the tree branches, the greater will be the deviation between the greedy and optimal tree errors. However, for signals in Besov space, this effect will be limited by the swift decay of the coefficients across scale.

## 4.3. Hierarchical wavelet representations

Like the linear and scalar nonlinear approximants, tree approximants induce (tree-based) hierarchical wavelet representations.

### 4.3.1. Greedy tree representation

Following ref,[18] normalize the maximum wavelet coefficient to magnitude 1 and define $\mathcal{T}_r$ to be the smallest tree containing all wavelet coefficients of magnitude $\geq 2^{-r+1}$, $r \geq 1$, with $\mathcal{T}_0$ containing the scaling coefficients. Define

$$\Sigma_0 := \sum_k u_{j_0,k} \phi_{j_0,k}, \tag{15}$$

$$\Sigma_r := \Sigma_0 + \sum_{I \in \mathcal{T}_r} w_I \psi_I, \qquad r = 1, 2, \ldots \tag{16}$$

$\Sigma_r$ corresponds to the greedy tree approximant $\widehat{f}_{\mathsf{G}}$ built from all wavelet terms $w_I \psi_I$ whose coefficients $|w_I| \geq 2^{-r}$; $\mathcal{T}_r$ is the corresponding greedy tree.

Define the difference tree approximants

$$\Delta_0 := \Sigma_0, \tag{17}$$

$$\Delta_r := \Sigma_r - \Sigma_{r-1}. \tag{18}$$

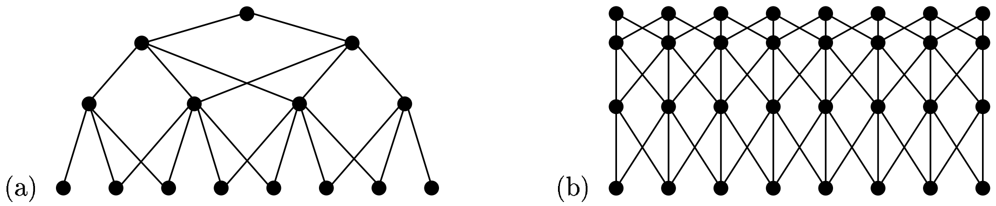


**Figure 4.** *Examples of more general tree structures for use with (a) longer wavelets than the Haar and (b) redundant wavelet transforms.*

Each $\Delta_r$ contains the wavelet terms $w_I \psi_I$ whose coefficients lie in the range $2^{-r} \leq |w_I| < 2^{-(r-1)}$.

Using these definitions, Cohen et al rewrite (2) as[18]

$$f = \sum_{r \geq 0} \Delta_r. \tag{19}$$

This decomposition is hierarchical in three different ways: *(space)* the basis atoms of the wavelet transform are multiscale in space; *(coefficients)* the $\Delta_r$ group wavelet coefficients of roughly similar size, with this size decreasing exponentially with $r$; and *(geometry)* when terminated at any finite $r$, the resulting representation lives on a connected rooted wavelet subtree. In fact, the terminated decomposition is a valid greedy tree approximation of $f$. This natural way to group wavelet coefficients has proven very useful for image compression.[18,19,24,25]

### 4.3.2. Optimal tree representation

Optimal tree approximation induces a new hierarchical wavelet representation. We merely redefine the $\Sigma_r$, $r \geq 1$, to be an optimal tree approximant with $L_2$ error between $2^{-r+1}\|f\|_2^2$ and $2^{-r}\|f\|_2^2$ (corresponding to the performance measure (12) lying between $(1 - 2^{-r})\|f\|_2^2$ and $(1 - 2^{-(r-1)})\|f\|_2^2$). Then, with $\Delta_r$ as in (17), (18), we have the representation (19).

Now, rather than grouping wavelet coefficients of roughly similar size, the $\Delta_r$ group coefficients in terms of those that contribute most to decreasing the tree approximation error. Furthermore, when terminated at any finite $r$, the resulting representation is an optimal tree approximation of $f$. This is also a very natural way to write the wavelet decomposition.

## 4.4. Extensions

When $j_0 > 0$, we have several coarsest-scale wavelet coefficients, with each rooting one of a forest of trees. While we could build optimal tree approximations for each tree separately, it is more natural to form one tree by tying the root wavelet coefficients together at a "super root." The same approach applies to higher-dimensional wavelet transforms. For instance, tensor image wavelet transforms have separate forests for each of the three subbands corresponding to the three wavelet orientations (vertical, horizontal, and diagonal).[1]

Less obviously, there is no reason to insist on a binary tree structure for modeling the persistence of large and small wavelet coefficient values. While the binary tree is matched to the Haar wavelet, for longer overlapping wavelets, trees with more than two children could prove useful (see Figure 4(a), for example). In this case, each tree node will have more than one parent. The key to all of this, of course, is to make the tree general enough that it can capture the time-frequency structure of the singularities in the data under consideration. It is not clear whether the $O(N)$ Solution 2 of Section 4.2 is applicable to approximation on such trees, since it assumes a nonoverlapping partitioning of the data. Solution 1 remains valid provided we take into account the multiplicity of parents.

Tree structures can also be placed on redundant wavelet transforms,[1,2] which retain the same number of wavelet coefficients at each scale (see Figure 4(b), for example), complex wavelet transforms,[28] and steerable pyramid transforms.[29]

## 5. APPLICATION TO SIGNAL AND IMAGE ESTIMATION

While optimal tree approximations could potentially have application in data compression (where they could help minimize the encoding distortion), they are immediately applicable to noise removal problems in signal and image processing.[20–22] In *denoising* we observe data $f$ that consists of a desired signal $g$ corrupted by additive white Gaussian noise. Translated into the wavelet domain, the problem reads:

$$\text{given} \;\; w_I = d_I + \eta_I \;\forall\, I, \qquad \text{estimate} \;\; d_I, \tag{20}$$

with $\{w_I\}$ and $\{d_I\}$ the wavelet transforms of $f$ and $g$, respectively, and $\{\eta_I\}$ a white Gaussian noise process. Traditionally we work to minimize the mean-square-error of the estimate.

A wavelet thresholding approach to denoising is simple to motivate. Assuming that most of the energy in the desired signal $g$ compacts into $n$ (large) wavelet coefficients, thresholding the noisy wavelet coefficients $\{w_I\}$ (that is, computing the $n$-term approximation $\{\widehat{w}_I\}$) will "keep" the signal and "kill" most of the noise (which is distributed uniformly in the wavelet domain). That is, $\{\widehat{w}_I\}$ should be a good estimator for the noise-free wavelet coefficients $\{d_I\}$.

Let $\mathcal{S}$ denote a (not necessarily connected) set of wavelet indices $I$ and set

$$\widehat{w}_I = \begin{cases} w_I, & I \in \mathcal{S} \quad \text{(keep)} \\ 0, & I \notin \mathcal{S} \quad \text{(kill).} \end{cases} \tag{21}$$

Define the best $\mathcal{S}$ as that solving the regularized optimization

$$\max_{\mathcal{S}} \sum_{I \in \mathcal{S}} |\widehat{w}_I|^2 - \mu |\mathcal{S}| \tag{22}$$

with $\mu > 0$. The parameter $\mu$ balances minimization of bias (by matching $\{\widehat{w}_I\}$ to $\{w_I\}$) with minimization of variance (by penalizing the complexity of the estimate $\mathcal{S}$).[23,20,21]** Different $\mu$ lead to different $|\mathcal{S}|$ and thus different thresholds. Donoho has derived threshold values that yield asymptotically optimal estimators for signals in Besov spaces.[23] See Figure 5(a)–(c) for an example with a simple test signal.††

One problem with threshold-based denoising is that occasionally the random fluctuations of the noise can cause "small" wavelet coefficients far down the wavelet tree to jump above the threshold, becoming "large." These spurious values are passed into the signal estimate as out-of-place wavelet atoms (see Figure 5(c)). This begs the question "which bumps are real bumps?"[23]

For many signals (in particular those fitting a piecewise smooth signal model), we can answer that the "real" bumps are those that arise from coherent persistence structures across scale. Tree approximation allows us to factor these structures into a wavelet thresholding denoising algorithm.

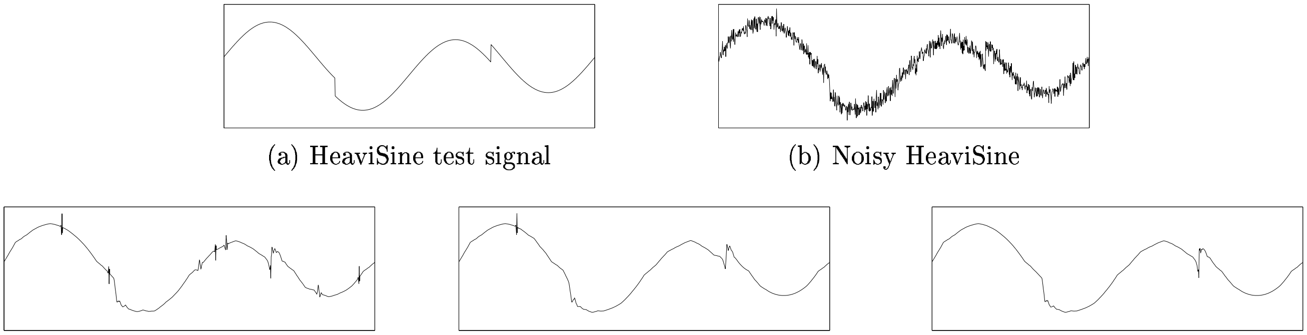
In tree denoising, we constrain the set $\mathcal{S}$ in (22) to be a connected, rooted subtree $\mathcal{T}$ (just as in tree approximation)..[20–22] This provides an extra regularization in addition to the complexity penalty: With the tree constraint, a wavelet coefficient cannot be considered "large" and included in the estimate unless all of its ancestors are "large." Such a thresholding pattern is fully capable of capturing both smooth signal structures and edges, but not spurious noise bumps (see Figure 5(e)). The tree approximation can of course be computed using either algorithm from Section 4.2. Donoho has derived thresholds for use with the Haar wavelet and certain smoothness spaces that provide asymptotic error optimality.[20,21]

Note that a signal estimate based on the greedy tree approximation of Section 4.1 does not have the same power to suppress spurious noise bumps, since it is implicitly based on scalar thresholding (see Figure 5(d)).

In Figure 6 we conduct a quantitative performance comparison of scalar, greedy tree, and optimal tree thresholding for denoising the test signal. Two features are evident: (1) both tree estimates boast a lower mean-square-error than the scalar estimate; and (2) the minima of both tree estimates occur to the right of the minimum of the scalar estimate, which indicates that at their optimal operating points, the tree estimates will contain more signal structure.

**See[17] for a derivation of soft threshold denoising as a least squares problem with a Besov norm regularization penalty.

††The random number generator was intialized with a seed obtained from a fortune cookie at Bloh Chau Chinese Restaurant, Houston.



(a) HeaviSine test signal

(b) Noisy HeaviSine

(c) Hard threshold estimate

(d) Greedy tree threshold estimate

(e) Optimal tree threshold estimate

**Figure 5.** *Signal denoising using approximation concepts. (a) HeaviSine test signal, (b) noisy test signal. Both the (c) scalar hard threshold estimate (mean-squared-error, MSE=39) and (d) greedy tree estimate (MSE=25) contain a large noise bump caused by a fine-scale wavelet coefficient creeping above the threshold. In contrast, since the (d) optimal tree estimate (MSE=21) naturally matches the structure of the singularities in this signal (recall Figure 1(d)), it completely suppresses this and other bumps. Details: 1024-point signal with maximum value=4, additive white Gaussian noise* $\sigma=\frac{1}{2}$*, Daubechies-6 wavelet, 6 scales,* $n = 18$ *terms in estimates.*

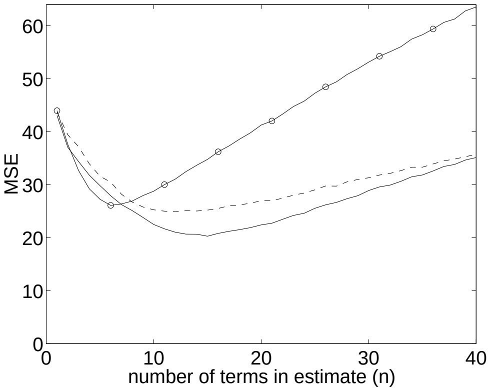


**Figure 6.** *MSEs for the experiment of Figure 5 for various* $n$*: (○) scalar nonlinear, (dash) greedy tree, and (solid) optimal tree estimates.*

## 6. CONCLUSIONS

This paper has aimed to elucidate the links between greedy tree[18] and optimal tree approximation.[20,21] While Dohono's fast $O(N)$ algorithm for solving the tree approximation problem should be sufficient for most applications, there could be situations where the $O(N \log N)$ CSSA is more appropriate. Regardless of the computational algorithms involved, tree approximation can impart considerable useful structure into wavelet signal processing algorithms. For example, by incorporating a priori knowledge on the structure of singularities in the wavelet domain, tree-based denoising suppresses spurious noise bumps. This scheme can thus be interpreted as a deterministic counterpart to the Bayesian statistical estimation scheme of ref.[11]

It is not clear that the optimal tree approach is directly applicable in the tree encoding scheme of Cohen et al,[18] because supernodes containing many small wavelet coefficient values can hide large values that eventually need to be coded (see Figure 2(e), for example). Moreover, given the rapid decay of the wavelet coefficients and the shallow depths of the wavelet trees used in practice, the difference between the greedy and optimal tree approximations may be only slight (recall Figure 3). An optimal tree encoder could potentially outperform a greedy tree in a perceptual metric, however (recall Figure 5).

**Acknowledgements:** Thanks to Ron DeVore for exposing me to greedy tree approximation and for many fruitful discussions. Thanks to David Donoho for pointing me to refs.[20–22]

## REFERENCES

1. S. Mallat, *A Wavelet Tour of Signal Processing.* Academic Press, 1998.
2. C. S. Burrus, R. A. Gopinath, and H. Guo, *Introduction to Wavelets and Wavelet Transforms: A Primer.* Prentice-Hall, 1998.
3. E. H. Adelson, "Layered representations for image coding," tech. rep., no. 181 MIT Media Lab, 1991.
4. U. Grenander, *Elements of Pattern Theory.* Baltimore: Johns Hopkins University Press, 1996.
5. K. S. Fu, "A step towards unification of syntactic and statistical pattern recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 5, pp. 200–205, 1983.
6. K. E. Mark, M. I. Miller, and U. Grenander, "Constrained stochastic language models," in *Image Models (and Their Speech Model Cousins)* (S. E. Levinson and L. Shepp, eds.), (Minneapolis), IMA Volumes in Mathematics and its Applications, 1994.
7. M. Basseville, A. Benveniste, K. C. Chou, S. A. Golden, R. Nikoukhah, and A. S. Willsky, "Modeling and estimation of multiresolution stochastic processes," *IEEE Trans. Inform. Theory*, vol. 38, pp. 766–784, Mar. 1992.
8. C. Bouman and M. Shapiro, "A multiscale random field model for Bayesian image segmentation," *IEEE Trans. Image Processing*, vol. 3, pp. 162–177, March 1994.
9. S. LoPresto, K. Ramchandran, and M. T. Orchard, "Image coding based on mixture modeling of wavelet coefficients and a fast estimation-quantization framework," in *Data Compression Conference '97*, (Snowbird, Utah), pp. 221–230, 1997.
10. E. Simoncelli, "Statistical models for images: Compression, restoration, and synthesis," in *Proc. 31st Asilomar Conf. on Signals, Systems, and Computers*, (Pacific Grove, CA), pp. 673–678, Nov. 1997.
11. M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 46, Apr. 1998. (Special Issue on Wavelets and Filter Banks).
12. K. E. Timmermann and R. D. Nowak, "Multiscale modeling and estimation of poisson processes with application to photon-limited imaging," *IEEE Trans. Inform. Theory*, vol. 45, pp. 846–862, Apr. 1999.
13. R. Nowak, "Multiscale hidden Markov models for Bayesian image analysis," in *Bayesian Inference in Wavelet Based Models* (B. Vidakovic and P. Müller, eds.), Lecture Notes in Statistics 141, Springer-Verlag, 1999.
14. R. Riedi, M. S. Crouse, V. J. Ribeiro, and R. G. Baraniuk, "A multifractal wavelet model with application to TCP network traffic," *IEEE Trans. Inform. Theory*, vol. 45, pp. 992–1018, Apr. 1999.
15. J. Scargle, "Studies in astronomical time series analysis. V. Bayesian blocks, a new method to analyze structure in photon counting data," *Astrophysical J.*, vol. 504, pp. 405–418, 1998.
16. R. A. DeVore, "Nonlinear approximation," *Acta Numerica*, pp. 51–150, 1998.
17. A. Chambolle, R. A. DeVore, N.-Y. Lee, and B. J. Lucier, "Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage," *IEEE Trans. Image Processing*, vol. 7, pp. 319–355, July 1998.
18. A. Cohen, W. Dahmen, I. Daubechies, and R. DeVore, "Tree approximation and encoding," 1999. Preprint.
19. D. L. Donoho, M. Vetterli, R. A. DeVore, and I. Daubechies, "Data compression and harmonic analysis," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2435–2476, Oct. 1998.
20. D. L. Donoho, "CART and best-ortho-basis: A connection," *Ann. Statist.*, vol. 25, no. 5, pp. 1870–1911, 1997.
21. D. L. Donoho, N. Dyn, D. Levin, and T. Yu, "Smooth multiwavelet duals of Alpert bases by moment-interpolating refinement," Dec. 1994.

22. J. Engel, “A simple wavelet approach to nonparametric regression from recursive partitioning schemes,” *J. Multivariate Anal.*, vol. 49, pp. 242–254, 1994.
23. D. L. Donoho, “De-noising by soft-thresholding,” *IEEE Trans. Inform. Theory*, vol. 41, pp. 613–627, May 1995.
24. J. Shapiro, “Embedded image coding using zerotrees of wavelet coefficients,” *IEEE Trans. Signal Processing*, vol. 41, pp. 3445–3462, Dec. 1993.
25. A. Said and W. 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.
26. R. G. Baraniuk, *Shear Madness: Signal-Dependent and Metaplectic Time-Frequency Representations.* Ph.D. dissertation, Dep. Elec. Comput. Eng., Univ. Illinois at Urbana-Champaign, 1992.
27. R. G. Baraniuk and D. L. Jones, “A signal-dependent time-frequency representation: Fast algorithm for optimal kernel design,” *IEEE Trans. Signal Processing*, vol. 42, pp. 134–146, Jan. 1994.
28. N. Kingsbury, “The dual-tree complex wavelet transform: A new technique for shift invariant and directional filters,” in *DSP Workshop ’98*, (Bryce Canyon, UT), 1998.
29. E. Simoncelli and H. Farid, “Steerable wedge filters for local orientation analysis,” *IEEE Trans. Image Processing*, vol. 9, pp. 1377–1382, Sept. 1996.