



Haar-Like Wavelets on Hierarchical Trees

Rick Archibald¹ · Ben Whitney²

Received: 1 August 2022 / Revised: 7 December 2023 / Accepted: 20 January 2024
© Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract

Discrete wavelet methods, originally formulated in the setting of regularly sampled signals, can be adapted to data defined on a point cloud if some multiresolution structure is imposed on the cloud. A wide variety of hierarchical clustering algorithms can be used for this purpose, and the multiresolution structure obtained can be encoded by a *hierarchical tree* of subsets of the cloud. Prior work introduced the use of *Haar-like bases* defined with respect to such trees for approximation and learning tasks on unstructured data. This paper builds on that work in two directions. First, we present an algorithm for constructing Haar-like bases on general discrete hierarchical trees. Second, with an eye towards data compression, we present thresholding techniques for data defined on a point cloud with error controlled in the L^∞ norm and in a Hölder-type norm. In a concluding trio of numerical examples, we apply our methods to compress a point cloud dataset, study the tightness of the L^∞ error bound, and use thresholding to identify MNIST classifiers with good generalizability.

Keywords Unstructured data · Lossy compression · Euclidean metric approximation

Mathematics Subject Classification 65T60

1 Introduction

A basic step common to many compression methods is the identification of some underlying structure in the data to be compressed. The structure should ideally be both simple and explanatory, requiring relatively few bits to encode and capturing most of the variation in the data. If deviations from the aforementioned structure can furthermore be encoded at low cost (or, in applications where lossy compression is acceptable, discarded without incurring too much error), then compression can be achieved.

✉ Rick Archibald
archibaldrk@ornl.gov
Ben Whitney
whitnebe@uwec.edu

¹ Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

² Mathematics Department, University of Wisconsin Eau Claire, Eau Claire, WI 54701, USA

23 The structure exploited by general-purpose compression methods is often repetition. Run-
24 length encoding [1–3] replaces repeated sequences of a token with one instance of the token
25 and a count. Dictionary methods [4–6] capture redundancies of longer strings, like words.
26 At a less granular level, deduplication techniques [7–9] can be used to eliminate repeated file
27 segments and filesystem blocks.

28 Scientific data, stored in high-precision formats, generally exhibit little byte-level redun-
29 dancy, and so compression methods tailored to scientific data typically make use of more
30 mathematically sophisticated structure. The case of data defined on regular grids has received
31 the most study, in part because well-developed image and video compression methods can
32 be adapted to that setting. Transform-based techniques (Fourier methods, wavelet methods,
33 and the like) are well-suited to such data, and a wide variety have been proposed [10–16].
34 Whereas a typical transform-based method decomposes its input using a predetermined basis
35 or filter bank, tensor methods, which are especially suited to high-dimensional data, use ten-
36 sor factors learned from the data [17–19]. Another common approach is to predict each input
37 value from the data already encoded and compress the residuals [20–25].

38 Wavelet-type techniques can be adapted to scientific data not laid out on a regular grid [26].
39 The main challenge is in appropriately defining the wavelet basis, which can then be used
40 for compression, denoising, etc. as usual. For graphical data, the basis is often constructed
41 by way of the graph Laplacian [27–30]. For data without graph network based structure, a
42 common approach is to apply a hierarchical clustering algorithm to the data and then define
43 the basis in terms of the tree structure obtained [31–34]. This technique is exemplified by the
44 2010 paper of Gavish et al., which proposes the use of *Haar-like bases* to learn from data
45 organized into balanced hierarchical trees. Gavish et al. relate data smoothness to the decay
46 rate of Haar-like basis coefficients and propose methods for function approximation, with
47 error controlled in the L^1 norm, and semisupervised learning, with expected error controlled
48 in the L^2 norm. These mathematically derived error bounds let scientists apply the methods
49 to their data with confidence.

50 This paper extends the work of Gavish et al. in two directions. Section 2 concerns the
51 use of Haar-like bases for decomposition and compression of functions defined on point sets
52 where we assume some arbitrary hierarchical tree structure. We establish the setting and
53 give an algorithm for generating Haar-like bases in Sect. 2. Thresholding methods with error
54 controlled in the L^∞ norm and in a Hölder-type norm are presented in Sect. 3. These methods
55 incorporate no information about any metric structure undergirding the point set, instead
56 relying on the metric structure of the tree. In the case of data defined on a Euclidean space, the
57 tree metric is not equivalent to the Euclidean metric, causing measures of function smoothness
58 on the tree to depart from the corresponding measures on the underlying Euclidean space.
59 Thresholding methods with error controlled in the L^∞ norm and in a Hölder-type norm are
60 presented in Sect. 3.

61 The paper allows for reconstruction and compression of data sets with minimal restrictions
62 on data structure. We begin our discussion by focusing completely on the domain, restricting
63 data only to a point set and then we define how hierarchical tree structures can be used to
64 provide the minimal structure necessary to build a Haar-like basis in Sect. 2. We use the
65 definitions and properties of the Haar-like basis to develop thresholding method capable
66 of compression of data on point set with guaranties error controls on either L^∞ norms or
67 Hölder-type norms in Sect. 3. We conclude in Sect. 4 with three numerical examples. We use
68 the Appendices to proof a few missing pieces not handled in the Sections. Specifically, we
69 proof that tree distance function used in this work is a metric in Appendix A and B, and we
70 proof the bounds in Remark 4 in Appendix C.

2 Setting and Construction of Haar-Like Bases

We begin with giving a table of common definitions used in this paper, and there first mention, in Table 1.

In this section, we focus on the domain of the underlying function and in Sect. 3, the function is considered together for error thresholding. We start by giving some basic definitions about partitions on graphs, and define some tools and structures used to act and build these partitions. With some description we define a hierarchical tree partition and use this to structure to derive a Haar-like transformation of this given tree. Let Ω be some discrete collection of points. Each element of Ω may be, for example, a point in Euclidean space, a molecular configuration, or an image. Our objective is to decompose and compress functions defined on Ω . We will assume that the point set Ω is organized into a sequence of increasingly fine partitions, the formal requirements on which are given in Definition 1. Informally, we require that

- (a) the coarsest partition is $\{\Omega\}$, the trivial partition;
- (b) the partitions are nested: each set in each partition is contained in a single set in the next coarsest partition, and is split into a discrete union of sets in the next finest partition; and
- (c) if Ω is discrete, the finest partition is $\{\{x\} : x \in \Omega\}$, the partition of Ω into singletons.

Such a collection has a natural graph structure [35], with the partition sets the nodes and an edge connecting two nodes if one contains the other and they belong to consecutive partitions. There exists a path from each partition set to Ω , the only set in the coarsest partition, so the graph is connected. Furthermore, we argue that the graph is acyclic. Suppose the graph contains a cycle; let $\Omega_1, \Omega_2, \Omega_3, \dots, \Omega_1$ be its node sequence. Being adjacent, Ω_1 and Ω_2 belong to consecutive partitions, and one contains the other. Without loss of generality, suppose that $\Omega_1 \supseteq \Omega_2$. By definition, Ω_2 is only adjacent to sets in the next coarsest partition (the partition containing Ω_1) or the next finest partition. If Ω_3 is contained in the partition containing Ω_1 , it must be Ω_1 , since every other set in that partition is disjoint from Ω_2 . Ω_3 cannot be Ω_1 , though, since a cycle cannot contain the same edge twice. Ω_3 must therefore be a set in the next finest partition. Repeating this argument, the ‘cycle’ continues to finer

Table 1 Common definitions used in this paper and the location of their first mention

Symbol	First Ref	Definition
$\Omega = \Omega_{\text{root}}$	pg. 4	Ω is a collection of discrete data points. It is also known at the root set when considering partitions
Ω_i	pg. 4	Partition such that $\Omega_i \subseteq \Omega$ represented by the index i
I	pg. 4	Index set for the nodes of the tree, containing all edge information
$\text{depth}(\Omega_j)$	pg. 4	Number of edges in the path from Ω to Ω_j
$\text{parent}(\Omega_j)$	pg.4	$\Omega_i = \text{parent}(\Omega_j)$, iff $\Omega_j \subseteq \Omega_i$ and $\text{depth}(\Omega_i) = \text{depth}(\Omega_j) - 1$
$\text{children}(\Omega_j)$	pg.4	The set s.t. $\{\Omega_j \in N : \Omega_i = \text{parent}(\Omega_j)\}$
leaves	pg.5	The index set s.t. $\{j \in I : \text{children}(j) \text{ is empty}\}$
branches	pg.5	The index set of the interior nodes of the tree given by the set $I \setminus \text{leaves}$
ν	pg.6	$\nu(\Omega_i) = \Omega_i / \Omega $, normalized counting measure on Ω
$ \Omega $	pg.7	The number of points in the set

99 and finer partitions and never returns to the partition containing Ω_1 . The graph is therefore
 100 acyclic. In particular, it is a tree, or, with Ω designated the root, a rooted tree. We call such
 101 a tree a *hierarchical tree*. A formal definition is given below.

102 Let Ω_j be a set in one of the partitions, and consider the path in the tree from Ω , the root,
 103 to Ω_j . Let $\text{depth}(\Omega_j)$ denote the number of edges in this path. Each set along the path
 104 contains Ω_j , and the sets become progressively smaller as the depth increases towards Ω_j .
 105 The collections of sets encountered along such paths will play a prominent role in this paper,
 106 so we introduce notation to simplify referring to them. Write $\Omega_i \preceq \Omega_j$ if the path from Ω to
 107 Ω_j is contained in the path from Ω to Ω_i . Using this notation, if N is the set of nodes of the
 108 tree, then $\{\Omega_i \in N : \Omega_i \preceq \Omega_j\}$ is the set of nodes encountered in the path from Ω to Ω_j .
 109 The set Ω_i in this path satisfying $\text{depth}(\Omega_i) = \text{depth}(\Omega_j) - 1$ is called the *parent* of Ω_j .
 110 (The root Ω has no parent.) We write $\Omega_i = \text{parent}(\Omega_j)$. We denote by $\text{children}(\Omega_i)$
 111 the set $\{\Omega_j \in N : \Omega_i = \text{parent}(\Omega_j)\}$.

112 For the purpose of referring to various objects associated with the sets in the partitions, it
 113 will be convenient to adopt an index set I for the nodes of the tree, writing $N = \{\Omega_i : i \in I\}$.
 114 Please see Fig. 1 for a simple example of the index set. We will, in a slight abuse of notation,
 115 substitute these indices into the notation introduced in the previous paragraph. That is, we
 116 will write

$$\begin{aligned}
 & i \leq j && \text{if } \Omega_i \preceq \Omega_j, \\
 & i = \text{parent}(j) && \text{if } \Omega_i = \text{parent}(\Omega_j), \\
 & \text{children}(i) && \text{for } \{j \in I : i = \text{parent}(j)\}, \text{ and} \\
 & \text{depth}(i) && \text{for } \text{depth}(\Omega_i).
 \end{aligned}$$

118 We will additionally denote by leaves the set $\{j \in I : \text{children}(j) \text{ is empty}\}$. Observe
 119 that $j \in \text{leaves}$ iff Ω_j is a leaf of the tree. The index set of the interior nodes of the tree
 120 are defined as branches which is the set $I \setminus \text{leaves}$. Figure 1 gives an illustration of this
 121 notation convention, where we note that the index set I additionally gives information about
 122 the path of each node. We can now give a formal definition of a hierarchical tree.

123 **Definition 1 (Hierarchical Tree)** Let $\{\Omega_i : i \in I\}$ be a collection of subsets of some set Ω .
 124 We call a rooted tree $(\{\Omega_i : i \in I\}, E, \Omega_{\text{root}})$ a *hierarchical tree* if

- 125 (a) $\Omega_{\text{root}} = \Omega$;
- 126 (b) for all $i \in \text{branches}$, $\Omega_i = \cup_{j \in \text{children}(i)} \Omega_j$; and
- 127 (c) for all distinct $x, y \in \Omega$, there exists $i \in I$ such that $\Omega_i \ni x$ and $\Omega_i \not\ni y$.

128 If additionally $0 < |\Omega_i| < \infty$ for all $i \in I$, we call the tree a *discrete* hierarchical tree.

129 All of the hierarchical trees considered in this paper will be discrete. We list the discreteness
 130 condition separately because most of the properties of hierarchical trees that we use do not
 131 depend on it.

132 See Fig. 1 for an illustration of an example discrete hierarchical tree. Observe that the
 133 leaf sets, each containing a single point, can have different depths. As a result, the sets of
 134 a particular depth do not necessarily form a partition of Ω in its entirety. In this regard,
 135 Definition 1 departs slightly from the informal description of a hierarchical tree given at the
 136 start of this Sect. 2.

137 In Fig. 1, each point x of Ω is the sole member of some (unique) leaf set Ω_j . Conversely,
 138 each leaf set Ω_j contains only a single point x of Ω . That is, the points of Ω are in bijection
 139 with the leaves of $\{\Omega_i : i \in I\}$, and so we can identify each point x of Ω by the leaf set Ω_j
 140 containing it. In Fig. 1, for example, $x_{(0,1)}$ is the point contained by $\Omega_{(0,1)}$. This identification
 141 can in fact be made in any discrete hierarchical tree, as shown by the following result.

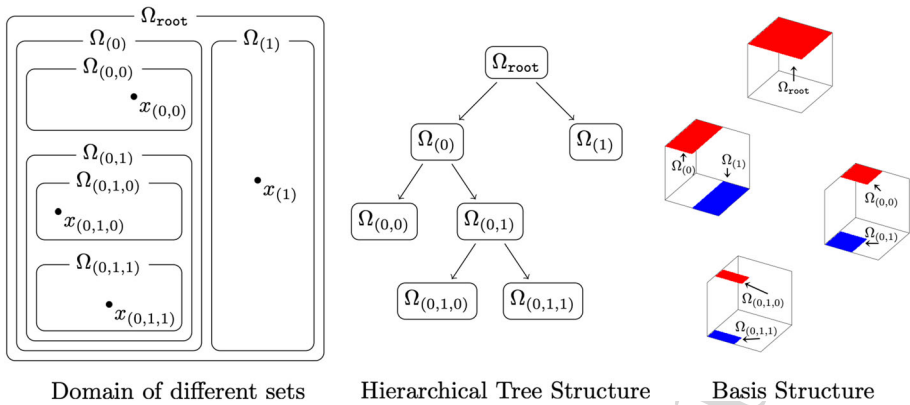


Fig. 1 Left: an illustration of a sequence of nested partitions of a set Ω comprising four points $x_{(0,1,0)}$, $x_{(0,1,1)}$, $x_{(0,0)}$, and $x_{(1)}$. The underlying index set I is $\{\text{root}, (0), (1), (0, 0), (0, 1), (0, 1, 0), (0, 1, 1)\}$. Note, the leaf sets of this tree are $\Omega_{(1)}$, $\Omega_{(0,0)}$, $\Omega_{(0,1,0)}$, and $\Omega_{(0,1,1)}$. The branches sets are Ω_{root} , $\Omega_{(0)}$ and $\Omega_{(0,1)}$. Center: an illustration of the hierarchical tree identified with the partitions. Right: an illustration of the Haar-like wavelets on the hierarchical tree, where detail construction of this basis is given in Algorithm 1

Lemma 1 Let $\{\Omega_i : i \in I\}$ be a discrete hierarchical tree. For $i \in \text{leaves}$, let x_i denote the sole member of Ω_i . The mapping $\text{leaves} \rightarrow \Omega$ given by $i \mapsto x_i$ is well-defined and bijective.

Proof The proof a set of Lemma's in the Appendices that are need for this proof. Appendix B, Lemma 10 implies that the mapping is well-defined, i.e., that for all $i \in \text{leaves}$ there exists some $x \in \Omega$ such that $\Omega_i = \{x\}$. Appendix B, Lemma 11 then implies surjectivity. It remains to show injectivity. Suppose there exist $i, i' \in \text{leaves}$ such that $x_i = x_{i'}$. Since $\Omega_i = \{x_i\}$ and $\Omega_{i'} = \{x_{i'}\}$ by Lemma 10, $\Omega_i \supseteq \Omega_{i'}$ and $\Omega_i \subseteq \Omega_{i'}$. Then $i \preceq i'$ and $i \succeq i'$ by Appendix A, Lemma 7(a), and so $i = i'$ by Appendix A, Lemma 5. \square

Functions defined on the point set Ω do not, in general, have any particular relationship with a given hierarchical tree $\{\Omega_i : i \in I\}$ on Ω . We can, however, use the tree structure as an aid in decomposing and compressing such functions. To do this, we define an inner product on the space of functions defined on Ω and then construct, using the hierarchical tree, a basis orthonormal with respect to that inner product. We begin with the necessary definitions.

Definition 2 (Haar-Like Basis) Let $\{\Omega_i : i \in I\}$ be a discrete hierarchical tree. Let $\nu : \{\Omega_i : i \in I\} \rightarrow [0, 1]$ be the normalized counting measure on Ω : $\nu(\Omega_i) = |\Omega_i| / |\Omega|$. Define the balance bounds \underline{B} and \overline{B} to be the minimum and maximum, respectively, of $\{\nu(\Omega_j) / \nu(\Omega_i) : i \in I \text{ and } j \in \text{children}(i)\}$.

Let V be the space of real-valued functions on Ω , and define an inner product $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{R}$ on V by $\langle f, g \rangle = \int_{\Omega} fg \, d\nu \equiv \frac{1}{|\Omega|} \sum_{x \in \Omega} f(x)g(x)$. We say that a subset \mathcal{B} of V is a Haar-like basis with respect to the tree if

- (a) \mathcal{B} comprises $\mathbf{1}_{\Omega}$ and, for each $i \in \text{branches}$, $|\text{children}(i)| - 1$ functions $\psi_{i,m}$ that are (i) supported on Ω_i and (ii) constant on each of the sets $\{\Omega_j : j \in \text{children}(i)\}$; and
- (b) \mathcal{B} is orthonormal with respect to $\langle \cdot, \cdot \rangle$.

We call the functions $\psi_{i,m}$ wavelets to distinguish them from the constant function $\mathbf{1}_{\Omega}$ which is the indicator function for the set Ω .

169 The only difference between Definition 2 and the corresponding definitions in [36] is
 170 that the underlying hierarchical tree here is slightly more general. That any set \mathcal{B} satisfying
 171 the conditions given in Definition 2 is in fact a basis follows from orthonormality and the
 172 following counting argument, whereas Gavish et. al. is derived only for balanced trees which
 173 we describe in the next paragraph as to restrictive. We note that this generality permits
 174 vanishing moments and two-scale relationships, given added restrictions to the underlying
 175 structure of the dataset and construction. We proceed using a Haar-like setup, and note that
 176 besides the generalization of Definition 2, that the other major advancement compared to
 177 [36], comes in Sect. 3 with our ability to control errors in Quantities of Interests (QoI) with
 178 our thresholding techniques. Let N and E denote the sets of nodes and edges, respectively,
 179 of the hierarchical tree. By Lemma 1, $|\Omega| = |\text{leaves}|$, and so

$$180 \quad \dim(V) = |\Omega| = |\text{leaves}| = |I| - |\text{branches}| = |N| - |\text{branches}|. \quad (1)$$

181 Because (N, E) is a tree, $|N| = 1 + |E|$ [35]. Using the definition of children and the
 182 fact that $|\text{children}(i)| = 0$ for any $i \in \text{leaves}$,

$$183 \quad |E| = \sum_{i \in I} |\text{children}(i)| = \sum_{i \in \text{branches}} |\text{children}(i)|. \quad (2)$$

184 Combining Eqs. 1 and 2, we find that

$$185 \quad \begin{aligned} \dim(V) &= 1 + \sum_{i \in \text{branches}} |\text{children}(i)| - |\text{branches}| \\ 186 \quad &= 1 + \sum_{i \in \text{branches}} (|\text{children}(i)| - 1) = |\mathcal{B}|. \end{aligned}$$

187 \mathcal{B} is therefore a basis for V .

188 A function $f \in V$ can be decomposed using a given Haar-like basis by expressing the
 189 function as a linear combination of the basis functions. The coordinates of the function with
 190 respect to the basis encode the function, and the function can be compressed by compressing
 191 those coordinates. The coordinates can be compressed using techniques such as thresholding
 192 (addressed in the next Sect. 3), bit plane encoding [14], or zerotree coding [37]. In order to
 193 apply these methods, we must be able to construct Haar-like bases with respect to arbitrary
 194 discrete hierarchical trees. In certain simple cases, this can be done easily. For example, when
 195 $\{\Omega_i : i \in I\}$ is binary ($|\text{children}(i)| = 2$ for all $i \in \text{branches}$) and perfectly balanced
 196 ($\underline{B} = \overline{B} = \frac{1}{2}$), it is straightforward to generalize the construction of Haar wavelets on an
 197 interval to a Haar-like basis for V ; see Remark 1. In general, though, a more robust approach
 198 is required. Algorithm 1 defines a procedure that constructs a Haar-like basis without any
 199 restriction on the arbitrary or balance bounds of the hierarchical tree. The algorithm constructs
 200 a Haar-like basis by first adding the scaling function $\mathbf{1}_\Omega$ and then proceeds iteratively to
 201 calculate each Haar-like wavelet. The correctness of Algorithm 1 is established in Lemma 2.

203 **Lemma 2** *Let $\{\Omega_i : i \in I\}$ be a discrete hierarchical tree. The set \mathcal{B} constructed by Algorithm*
 204 *1 is a Haar-like basis for V .*

205 **Proof** We first show that the matrix M defined in Algorithm 1 Line 8 is unitary. This can be
 206 done by referring to Jarlskog’s parametrization of unitary matrices [38] and verifying that M
 207 has the specified form. We can also directly check that MM^T is the $n \times n$ identity matrix.

Algorithm 1 Construction of a Haar-like basis for V . I_{n-1} denotes the $(n - 1) \times (n - 1)$ identity matrix.

```

1: function CONSTRUCTBASIS(discrete hierarchical tree  $\{\Omega_i : i \in I\}$ )
2:   initialize  $\mathcal{B}$  to  $\{\}$ 
3:   add  $\mathbf{1}_\Omega$  to  $\mathcal{B}$ 
4:   for  $i$  in branches do
5:     fix an ordering  $j_1, \dots, j_n$  of children( $i$ )
6:     define  $\mathbf{w} \in \mathbb{R}^n$  by  $w_k = v(\Omega_{j_k})/v(\Omega_i)$ 
7:     define  $\mathbf{v} \in \mathbb{R}^{n-1}$  by  $v_k = \sqrt{w_k}$ 
8:     define  $M \in \mathbb{R}^{n \times n}$  by

```

$$M = \left[\begin{array}{c|c} I_{n-1} - \frac{1}{1+v_n} \mathbf{v}\mathbf{v}^\top & \mathbf{v} \\ \hline -\mathbf{v}^\top & v_n \end{array} \right]$$

```

9:   for  $m = 1, \dots, n - 1$  do
10:     define  $\psi_{i,m} \in V$  by

```

$$\psi_{i,m} = \sum_{k=1}^n \frac{M_{k,m}}{\sqrt{v(\Omega_{j_k})}} \mathbf{1}_{\Omega_{j_k}}$$

```

11:     add  $\psi_{i,m}$  to  $\mathcal{B}$ 
12:   end for
13:   end for
14:   return  $\mathcal{B}$ 
15: end function

```

208 Denote the blocks of M as follows:

209
$$M = \left[\begin{array}{c|c} I_{n-1} - \frac{1}{1+v_n} \mathbf{v}\mathbf{v}^\top & \mathbf{v} \\ \hline -\mathbf{v}^\top & v_n \end{array} \right] = \left[\begin{array}{c|c} A & B \\ \hline -B^\top & C \end{array} \right].$$

210 The product MM^\top can be written in terms of these blocks:

211
$$MM^\top = \left[\begin{array}{c|c} A & B \\ \hline -B^\top & C \end{array} \right] \left[\begin{array}{c|c} A^\top & -B \\ \hline B^\top & C^\top \end{array} \right] = \left[\begin{array}{c|c} AA^\top + BB^\top & -AB + BC^\top \\ \hline -B^\top A^\top + B^\top C & B^\top B + CC^\top \end{array} \right].$$

212 It suffices to show that $AA^\top + BB^\top = I_{n-1}$, $-AB + BC^\top = \mathbf{0}$, and $B^\top B + CC^\top = 1$. In
 213 brief,

214
$$AA^\top + BB^\top = \left(I_{n-1} - \frac{1}{1+v_n} \mathbf{v}\mathbf{v}^\top \right) \left(I_{n-1} - \frac{1}{1+v_n} \mathbf{v}\mathbf{v}^\top \right) + \mathbf{v}\mathbf{v}^\top$$

215
$$= I_{n-1}^2 + \frac{-2(1+v_n) + (1-v_n^2) + (1+v_n)^2}{(1+v_n)^2} \mathbf{v}\mathbf{v}^\top = I_{n-1},$$

216
$$-AB + BC^\top = -\left(I_{n-1} - \frac{1}{1+v_n} \mathbf{v}\mathbf{v}^\top \right) \mathbf{v} + v_n \mathbf{v}$$

217
$$= \left[\frac{1-v_n^2}{1+v_n} - (1-v_n) \right] \mathbf{v} = \mathbf{0}, \text{ and}$$

218
$$B^\top B + CC^\top = \mathbf{v}^\top \mathbf{v} + v_n^2 = \|\mathbf{v}\|^2 + v_n^2 = 1.$$

219 The key identity used here is $\|\mathbf{v}\|^2 + v_n^2 = 1$, which follows from the definition of \mathbf{v} in
 220 Algorithm 1 Line 7:

221
$$\|\mathbf{v}\|^2 + v_n^2 = \sum_{k=1}^n \sqrt{w_k}^2 = \sum_{k=1}^n \frac{|\Omega_{j_k}|}{|\Omega_i|} = 1.$$

We now check that the collection \mathcal{B} returned by Algorithm 1 satisfies the conditions given in Definition 2.

- (a) The inclusion $\mathbf{1}_\Omega \in \mathcal{B}$ is ensured by Algorithm 1 Line 3. Refer next to Algorithm 1 Line 9 - 12. Each $\psi_{i,m}$ is a linear combination of the indicator functions $\{\mathbf{1}_{\Omega_j} : j \in \text{children}(i)\}$. Because $\text{supp}(\mathbf{1}_{\Omega_j}) = \Omega_j \subseteq \Omega_i$, $\text{supp}(\psi_{i,m}) \subseteq \Omega_i$. Here, $\text{supp}(\mathbf{1}_{\Omega_j})$ is defined to be the support of the function $\mathbf{1}_{\Omega_j}$. Because $\{\Omega_j : j \in \text{children}(i)\}$ are disjoint, $\psi_{i,m}$ is constant on each Ω_j .
- (b) We claim that $\langle f, g \rangle = \delta_{f,g}$ for all $f, g \in \mathcal{B}$. There are three cases to consider.

1. $\langle \mathbf{1}_\Omega, \mathbf{1}_\Omega \rangle$: This case is immediate:

$$\langle \mathbf{1}_\Omega, \mathbf{1}_\Omega \rangle = \frac{1}{|\Omega|} \sum_{x \in \Omega} \mathbf{1}_\Omega(x) \mathbf{1}_\Omega(x) = \frac{|\Omega|}{|\Omega|} = 1.$$

2. $\langle \mathbf{1}_\Omega, \psi_{i,m} \rangle$: As $\psi_{i,m}$ is supported on Ω_i ,

$$\langle \mathbf{1}_\Omega, \psi_{i,m} \rangle = \frac{1}{|\Omega|} \sum_{x \in \Omega} \mathbf{1}_\Omega(x) \psi_{i,m}(x) = \frac{1}{|\Omega|} \sum_{x \in \Omega_i} \psi_{i,m}(x).$$

Let j_1, \dots, j_n be the ordering fixed in Algorithm 1 Line 5. Then $\psi_{i,m}$ takes value $M_{k,m}/\sqrt{v(\Omega_{j_k})}$ on Ω_{j_k} , and so

$$\begin{aligned} \langle \mathbf{1}_\Omega, \psi_{i,m} \rangle &= \frac{1}{|\Omega|} \sum_{k=1}^n |\Omega_{j_k}| \frac{M_{k,m}}{\sqrt{v(\Omega_{j_k})}} = \frac{1}{\sqrt{|\Omega|}} \sum_{k=1}^n \sqrt{|\Omega_{j_k}|} M_{k,m} \\ &= \frac{\sqrt{|\Omega_i|}}{\sqrt{|\Omega|}} \sum_{k=1}^n \frac{\sqrt{|\Omega_{j_k}|}}{\sqrt{|\Omega_i|}} M_{k,m} = \frac{\sqrt{|\Omega_i|}}{\sqrt{|\Omega|}} \sum_{k=1}^n M_{k,n} M_{k,m}. \end{aligned}$$

Since M has orthogonal columns and $m \neq n$, this equals zero.

- 3. $\langle \psi_{i,m}, \psi_{i',m'} \rangle$: If $i \not\leq i'$ and $i' \not\leq i$, then $\Omega_i = \text{supp}(\psi_{i,m})$ and $\Omega_{i'} = \text{supp}(\psi_{i',m'})$ are disjoint by Lemma 7(b) and so the inner product is zero. So, suppose, without loss of generality, that $i \leq i'$. If $i \neq i'$, then there exists some $j \in \text{children}(i)$ such that $j \leq i'$. Lemma 7(b) then implies that $\Omega_j \supseteq \Omega_{i'}$. Since ψ_i is constant on Ω_j , it is constant on $\Omega_{i'}$, and the result then follows from the previous case. If instead $i = i'$, let j_1, \dots, j_n be the (common) ordering fixed in Algorithm 1 Line 5. Then

$$\begin{aligned} \langle \psi_{i,m}, \psi_{i',m'} \rangle &= \frac{1}{|\Omega|} \sum_{x \in \Omega} \psi_{i,m}(x) \psi_{i',m'}(x) \\ &= \frac{1}{|\Omega|} \sum_{k=1}^n |\Omega_{j_k}| \frac{M_{k,m}}{\sqrt{v(\Omega_{j_k})}} \frac{M_{k,m'}}{\sqrt{v(\Omega_{j_k})}} \\ &= \frac{|\Omega|}{|\Omega|} \sum_{k=1}^n M_{k,m} M_{k,m'} = \delta_{m,m'}. \end{aligned}$$

□

We conclude this section with definitions of a metric on Ω and two norms on V . The function norms will be used in the next in Sect. 3 to measure the errors induced when compressing functions defined on Ω .

253 **Definition 3 (lca Norm)** Let $\{\Omega_i : i \in I\}$ be a hierarchical tree.

- 254 (a) Let $x, y \in \Omega$ be distinct. The set Ω_i of maximal depth satisfying $x, y \in \Omega_i$ is called the
 255 *lowest common ancestor* of x and y . We denote it $\text{lca}(x, y)$.
 256 (b) Let $j, j' \in I$. The index i of maximal depth satisfying $i \leq j, j'$ is called the *lowest*
 257 *common ancestor* of j and j' . We denote it $\text{lca}(j, j')$.

258 Proofs that lca is well-defined in each case can be found in Appendix A (see Lemma 6
 259 and 8).

260 **Definition 4 (Tree Metric and Hölder Seminorm)** Let $\{\Omega_i : i \in I\}$ be a discrete hierarchical
 261 tree. The *tree metric* $d : \Omega \times \Omega \rightarrow [0, 1]$ is defined by

$$262 \quad d(x, y) = \begin{cases} \nu(\text{lca}(x, y)) & x \neq y \\ 0 & \text{otherwise.} \end{cases}$$

263 Let $\|\cdot\|_{C^0} : V \rightarrow [0, \infty)$ denote the supremum norm: $\|f\|_{C^0} = \max_{x \in \Omega} |f(x)|$. For $\alpha \in$
 264 $(0, 1]$, we define the *Hölder seminorm* $|f|_{C^\alpha} : V \rightarrow [0, \infty)$ by

$$265 \quad |f|_{C^\alpha} = \max_{\substack{x, y \in \Omega \\ x \neq y}} \frac{|f(x) - f(y)|}{d(x, y)^\alpha}$$

266 and the *Hölder norm* $\|\cdot\|_{C^\alpha} : V \rightarrow [0, \infty)$ by $\|f\|_{C^\alpha} = \|f\|_{C^0} + |f|_{C^\alpha}$.

267 The definitions of the tree metric and the Hölder seminorm are taken from [36]. A proof
 268 that d is in fact a metric is given in Appendix B (see Lemma 9).

269 3 Thresholding Algorithms

270 This section concerns the problem of using the hierarchical structure of a point set to compress
 271 functions on that set so that the induced error is small in some error norm, specifically we wish
 272 to control the L^∞ norm and Hölder-type norms in our compression method. Our strategy,
 273 familiar from wavelet analysis, is to write the input function as a weighted sum of Haar-like
 274 basis functions and to then retain or discard each component according to its contribution
 275 (measured by the error norm) to the data. Given a function $f \in V$ and a Haar-like basis \mathcal{B}
 276 for V , we let \hat{f}_{DC} denote the coefficient $\langle f, \mathbf{1}_\Omega \rangle$ and $\hat{f}_{i,m}$ denote the coefficient $\langle f, \psi_{i,m} \rangle$.
 277 The task is then to decide whether to retain or discard each $\hat{f}_{i,m}$. We base our technique for
 278 thresholding with Hölder norm error control on the following result, proved in [36].¹

279 **Theorem 1** ([36, Theorem 2]) Let $\{\Omega_i : i \in I\}$ be a discrete hierarchical tree, and let \mathcal{B} be
 280 a Haar-like basis for V . Suppose $f \in V$ satisfies, for some $C \geq 0$ and $\alpha \in (0, 1]$,

$$281 \quad |\hat{f}_{i,m}| \leq C \nu(\Omega_i)^{\alpha+1/2}$$

282 for all wavelets $\psi_{i,m} \in \mathcal{B}$. Then

$$283 \quad |f|_{C^\alpha} \leq \frac{2C}{\underline{B}^{3/2}(1 - \overline{B}^\alpha)}.$$

¹ Theorem 1 differs slightly in the assumptions made of the hierarchical tree, but the proof of [36, Theorem 2] carries over with minimal modification.

284 Theorem 1 suggests a thresholding strategy to control L^∞ norm and Hölder-type norms:
 285 retain those coefficients $\hat{f}_{i,m}$ with $|\hat{f}_{i,m}| > C\nu(\Omega_i)^{\alpha+1/2}$ and discard the rest. The threshold-
 286 ing error will then be small as measured by the Hölder seminorm; to control the error in the full
 287 Hölder norm, we must additionally bound its supremum norm, since $\|\cdot\|_{C^\alpha} = \|\cdot\|_{C^0} + |\cdot|_{C^\alpha}$.
 288 In the next result, the supremum norm is bounded by expressing the values taken by f as
 289 deviations from the mean.

290 **Lemma 3** *Let $\{\Omega_i : i \in I\}$ be a discrete hierarchical tree. For all $f \in V$ and all $\alpha \in (0, 1]$,*
 291 *$\|f\|_{C^0} \leq |\hat{f}_{DC}| + |f|_{C^\alpha}$. As a result, $\|f\|_{C^\alpha} \leq |\hat{f}_{DC}| + 2|f|_{C^\alpha}$.*

292 **Proof** Fix $f \in V$. Observe that \hat{f}_{DC} is the mean of f :

293
$$\frac{1}{|\Omega|} \sum_{y \in \Omega} f(y) = \frac{1}{|\Omega|} \sum_{y \in \Omega} f(y) \mathbf{1}_\Omega(y) = \langle f, \mathbf{1}_\Omega \rangle = \hat{f}_{DC}.$$

294 For all $x \in \Omega$,

295
$$f(x) = \frac{1}{|\Omega|} \sum_{y \in \Omega} f(y) + \frac{1}{|\Omega|} \sum_{y \in \Omega} f(x) - f(y) = \hat{f}_{DC} + \frac{1}{|\Omega|} \sum_{y \in \Omega} f(x) - f(y).$$

296 By the definition of the Hölder seminorm, $|f(x) - f(y)| \leq |f|_{C^\alpha} d(x, y)^\alpha$. $d(x, y)$ is non-
 297 negative and at most one, so $|f|_{C^\alpha} d(x, y)^\alpha \leq |f|_{C^\alpha}$. As a result,

298
$$|f(x)| \leq |\hat{f}_{DC}| + \frac{1}{|\Omega|} \sum_{y \in \Omega} |f(x) - f(y)| \leq |\hat{f}_{DC}| + |f|_{C^\alpha}.$$

299 Take the maximum over $x \in \Omega$ to find that $\|f\|_{C^0} \leq |\hat{f}_{DC}| + |f|_{C^\alpha}$. Substituting into the
 300 definition of $\|f\|_{C^\alpha}$ then yields the final bound. □

301 We now use Lemma 3 to design a thresholding algorithm that induces limited errors in
 302 the Hölder norm.

303 **Theorem 2** *Let $\{\Omega_i : i \in I\}$ be a discrete hierarchical tree, and let \mathcal{B} be a Haar-like basis*
 304 *for V . Let $\alpha \in (0, 1]$ and $C \geq 0$. Given $f \in V$, define $\tilde{f} \in V$ by $\hat{f}_{DC}^\wedge = \hat{f}_{DC}$ and*

305
$$\tilde{f}_{i,m}^\wedge = \begin{cases} \hat{f}_{i,m} & |\hat{f}_{i,m}| > C\nu(\Omega_i)^{\alpha+1/2} \\ 0 & \text{otherwise} \end{cases}$$

306 for all wavelets $\psi_{i,m} \in \mathcal{B}$. Then

307
$$\|f - \tilde{f}\|_{C^\alpha} \leq \frac{4C}{\underline{B}^{3/2}(1 - \overline{B}^\alpha)}.$$

308 **Proof** By linearity,

309
$$(f - \tilde{f})_{i,m}^\wedge = \begin{cases} 0 & \hat{f}_{i,m} > C\nu(\Omega_i)^{\alpha+1/2} \\ \hat{f}_{i,m} & \text{otherwise} \end{cases}$$

310 for all wavelets $\psi_{i,m} \in \mathcal{B}$. In particular,

311
$$|(f - \tilde{f})_{i,m}^\wedge| \leq C\nu(\Omega_i)^{\alpha+1/2} \quad \text{and so} \quad |f - \tilde{f}|_{C^\alpha} \leq \frac{2C}{\underline{B}^{3/2}(1 - \overline{B}^\alpha)}$$

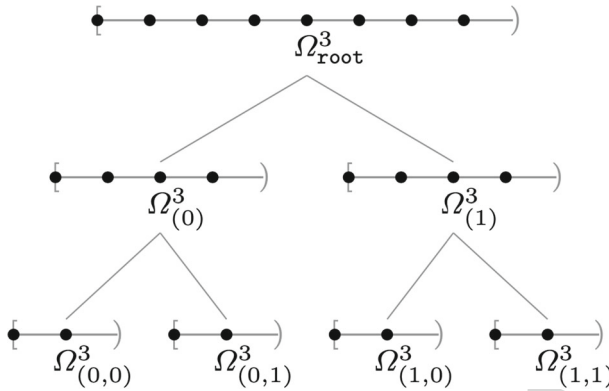


Fig. 2 The first three levels of $\{\Omega_b^3 : b \in B^3\}$. The root is $\Omega_{\text{root}}^3 = \Omega_0^3 = \Omega^3$; the leaves are not shown. Ω_b^3 contains x_c iff c starts with b . For example, $\Omega_{(0,1)}^3$ comprises two points, $x_{(0,1,0)} = \frac{1}{4}$ and $x_{(0,1,1)} = \frac{3}{8}$. As claimed, $|\Omega_{(0,1)}^3| = 2^{3-|(0,1)|} = 2$

312 by 1. Applying 3,

313
$$\|f - \tilde{f}\|_{C^\alpha} \leq \left| (f - \tilde{f})_{\text{DC}}^\wedge \right| + 2 \|f - \tilde{f}\|_{C^\alpha} \leq \frac{4C}{\underline{B}^{3/2}(1 - \overline{B}^\alpha)},$$

314 since $(f - \tilde{f})_{\text{DC}}^\wedge = 0$, again by linearity. □

315 What if we wish to bound the thresholding error in the supremum norm instead of the
 316 Hölder norm? The naïve approach of taking $\alpha \rightarrow 0$ in Theorem 2 is insufficient, as shown
 317 by Remark 1.

318 **Remark 1** We claim that no analogue of Theorem 2 holds when $\alpha = 0$, i.e., that the conditions
 319 $\hat{f}_{\text{DC}} = 0$ and $|\hat{f}_{i,m}| \leq \nu(\Omega_i)^{1/2}$ are insufficient to guarantee a bound on $\|f\|_{C^0}$ in terms only
 320 of \underline{B} and \overline{B} .² We demonstrate the impossibility of such a bound by constructing a family
 321 of hierarchical trees $\{\{\Omega_b^n : b \in B^n\} : n \in \mathbb{N}\}$ with uniform balance bounds, specifically
 322 $\underline{B} = \overline{B} = \frac{1}{2}$ and associated functions $f^n \in V^n$ satisfying $\hat{f}_{\text{DC}}^n = 0$ and $|\hat{f}_{b,m}^n| \leq \nu(\Omega_b^n)^{1/2}$
 323 such that $\|f^n\|_{C^0}$ grows unboundedly with the size of the tree (with n).

324 Take $n \in \mathbb{N}$ and let B^n be the set of binary sequences of length at most n . For $b, b' \in B^n$,
 325 denote by $|b|$ the length of b and by $b + b'$ the sequence obtained by concatenating b
 326 and b' . For $c \in B^n \setminus B^{n-1}$ (that is, c with length exactly n), define $x_c = \sum_{i=1}^n 2^{-i} c_i$.
 327 Define Ω^n to be the set $\{x_c : c \in B^n \setminus B^{n-1}\}$; observe that $|\Omega^n| = 2^n$. Given $b \in B^n$ and
 328 $c \in B^n \setminus B^{n-1}$, say that c starts with b if $b_i = c_i$ for all $1 \leq i \leq |b|$. For $b \in B^n$, define
 329 Ω_b^n to be the set $\{x_c : c \in B^n \setminus B^{n-1} \text{ and } c \text{ starts with } b\}$; observe that $|\Omega_b^n| = 2^{n-|b|}$, so that
 330 $|\Omega_b^n| / |\Omega^n| = 2^{-|b|}$. It is straightforward to verify that $\{\Omega_b^n : b \in B^n\}$ is a hierarchical tree
 331 on Ω^n and that $\underline{B} = \overline{B} = \frac{1}{2}$. $\{\Omega_b^3 : b \in B^3\}$ is depicted in Fig. 2.

332 Next we define a Haar-like basis for V^n . $\{\Omega_b^n : b \in B^n\}$ is by construction a binary
 333 tree, so the basis must contain a single wavelet $\psi_{b,1}^n$ for each $b \in \text{branches} = B^{n-1}$.
 334 For $b \in \text{branches}$, define $\psi_{b,1}^n = -2^{|b|/2} \mathbf{1}_{\Omega_{b+(0)}^n} + 2^{|b|/2} \mathbf{1}_{\Omega_{b+(1)}^n}$, and let \mathcal{B}^n be the basis

² The stipulation that the bound depend only on \underline{B} and \overline{B} , and not on size of the tree, is essential, and in fact a bound in terms of \underline{B} and the size of the tree always holds (take $a_i = 1$ in Theorem 3).

335 comprising $\mathbf{1}_{\Omega^n}$ and these wavelets $\psi_{b,1}^n$. $\mathbf{1}_{\Omega^n}$ has norm 1 because ν is a probability measure,
 336 and the same is true of the wavelets:

$$\begin{aligned}
 \langle \psi_{b,1}^n, \psi_{b,1}^n \rangle &= \int_{\Omega^n} [-2^{|b|/2} \mathbf{1}_{\Omega_{b+(0)}^n} + 2^{|b|/2} \mathbf{1}_{\Omega_{b+(1)}^n}]^2 \, d\nu \\
 &= 2^{|b|} [\nu(\Omega_{b+(0)}^n) + \nu(\Omega_{b+(1)}^n)] = 1.
 \end{aligned}$$

337
 338
 339 Next we show orthogonality. Suppose $\psi_{b,1}^n$ and $\psi_{b',1}^n$ are distinct wavelets with intersecting
 340 supports Ω_b^n and $\Omega_{b'}^n$, respectively. $\Omega_b^n \cap \Omega_{b'}^n \neq \emptyset$ implies $b \leq b'$ or $b' \leq b$ by Lemma 7(b).
 341 $b \neq b'$ by assumption; suppose $b' < b$ without loss of generality. $\psi_{b,1}^n$ is then supported on
 342 either $\Omega_{b'+(0)}^n$ or $\Omega_{b'+(1)}^n$, using Lemma 7(a). $\psi_{b',1}^n$ is constant on both of these sets, so it
 343 suffices to show that $\psi_{b,1}^n \perp \mathbf{1}_{\Omega^n}$:

$$\begin{aligned}
 \langle \psi_{b,1}^n, \mathbf{1}_{\Omega^n} \rangle &= \int_{\Omega^n} -2^{|b|/2} \mathbf{1}_{\Omega_{b+(0)}^n} + 2^{|b|/2} \mathbf{1}_{\Omega_{b+(1)}^n} \, d\nu \\
 &= 2^{|b|/2} [-\nu(\Omega_{b+(0)}^n) + \nu(\Omega_{b+(1)}^n)] = 0.
 \end{aligned}$$

344 We conclude that \mathcal{B}^n is a Haar-like basis. In fact for these uniform sets and balanced tree, the
 345 basis sets produce conventional Haar functions.

346 With the goal of developing a counter example to Theorem 2 when $\alpha \rightarrow 0$, we next
 347 we construct a function $f^n \in V^n$ such that $\hat{f}_{DC}^n = 0$ and $|\hat{f}_{b,1}^n| \leq \nu(\Omega_b^n)^{1/2}$ but $\|f^n\|_{C^0}$
 348 depends increasingly and unboundedly on n . Define $F: \{0, 1\} \rightarrow \{-1, 1\}$ by $F(0) = -1$
 349 and $F(1) = 1$. Define $f^n \in V^n$ by $f^n(x_c) = \sum_{i=1}^n F(c_i)$. The average of f^n is zero:

$$\hat{f}_{DC}^n = \frac{1}{|\Omega^n|} \sum_{x_c \in \Omega^n} \sum_{i=1}^n F(c_i) = \frac{1}{|\Omega^n|} \sum_{i=1}^n \sum_{x_c \in \Omega^n} F(c_i) = 0.$$

350 Now we turn to $\hat{f}_{b,1}^n$. Fix a wavelet $\psi_{b,1}^n$, and let B_b^n denote the set $B^{n-|b|-1} \setminus B^{n-|b|-2}$;
 351 observe that $|B_b^n| = 2^{n-|b|-1}$. $\psi_{b,1}^n$ is supported on the sets $\Omega_{b+(0)}^n = \{x_{b+(0)+c} : c \in B_b^n\}$
 352 and $\Omega_{b+(1)}^n = \{x_{b+(1)+c} : c \in B_b^n\}$. On these sets,

$$\begin{aligned}
 f^n(x_{b+(0)+c}) &= \sum_{i=1}^{|b|} F(b_i) - 1 + \sum_{i=1}^{n-|b|-1} F(c_i) \quad \text{and} \\
 f^n(x_{b+(1)+c}) &= \sum_{i=1}^{|b|} F(b_i) + 1 + \sum_{i=1}^{n-|b|-1} F(c_i).
 \end{aligned}$$

353 In particular, $f^n(x_{b+(1)+c}) - f^n(x_{b+(0)+c}) = 2$ for $c \in B_b^n$. As a result,

$$\begin{aligned}
 \langle f^n, \psi_{b,1}^n \rangle &= \int_{\Omega_{b+(0)}^n} -2^{|b|/2} f^n \, d\nu + \int_{\Omega_{b+(1)}^n} 2^{|b|/2} f^n \, d\nu \\
 &= \frac{1}{|\Omega^n|} \sum_{c \in B_b^n} -2^{|b|/2} f^n(x_{b+(0)+c}) + \frac{1}{|\Omega^n|} \sum_{c \in B_b^n} 2^{|b|/2} f^n(x_{b+(1)+c}) \\
 &= \frac{1}{|\Omega^n|} \sum_{c \in B_b^n} 2^{|b|/2} [f^n(x_{b+(1)+c}) - f^n(x_{b+(0)+c})] \\
 &= \frac{1}{|\Omega^n|} 2^{|b|/2+1} |B_b^n| = 2^{-n} 2^{|b|/2+1} 2^{n-|b|-1} = 2^{-|b|/2}.
 \end{aligned}$$

363 In particular, since $v(\Omega_b^n) = 2^{-|b|}$, $|\hat{f}_{b,1}^n| \leq v(\Omega_b^n)^{1/2}$. Nevertheless, $\|f^n\|_{C^0}$ cannot be
 364 bounded independent of n : $f^n(x_{(1,\dots,1)}) = \sum_{i=1}^n 1 = n$.

365 $|\hat{f}_{i,m}| \leq v(\Omega_i)^{1/2}$ is not, then, the correct decay rate for control of the supremum norm of
 366 f . A slightly faster decay in the wavelet coefficients is sufficient, as shown in the next result.

367 **Theorem 3** *Let $\{\Omega_i : i \in I\}$ be a discrete hierarchical tree, and let \mathcal{B} be a Haar-like basis*
 368 *for V . Let $\{a_i : i \in I\}$ be a collection of nonnegative numbers. If $f \in V$ satisfies*

369
$$|\hat{f}_{i,m}| \leq a_i v(\Omega_i)^{1/2}$$

370 *for all wavelets $\psi_{i,m}^n \in \mathcal{B}$, then*

371
$$\|f\|_{C^0} \leq |\hat{f}_{DC}| + \underline{B}^{-3/2} \max_{x_j \in \Omega} \sum_{i < j} a_i.$$

372 **Proof** We first state four bounds, all of which can be found in [36], for use in the proof. Let
 373 $i \in \text{branches}$.

374 (a) By the definition of \underline{B} , $v(\Omega_j) \geq \underline{B}v(\Omega_i)$ for all $j \in \text{children}(i)$.

375 (b) Applying this inequality, we can write

376
$$v(\Omega_i) = \sum_{j \in \text{children}(i)} v(\Omega_j) \geq \underline{B} |\text{children}(i)| v(\Omega_i),$$

377 so that $|\text{children}(i)| \leq \underline{B}^{-1}$. In particular, writing $\text{nc}(i) = |\text{children}(i)|$, $\text{nc}(i) -$
 378 $1 \leq \underline{B}^{-1}$.

379 (c) Let $\psi_{i,m}$ be any of the wavelets $\psi_{i,1}, \dots, \psi_{i,\text{nc}(i)-1}$. $\psi_{i,m}$ is constant on the sets $\{\Omega_j :$
 380 $j \in \text{children}(i)\}$; in a slight abuse of notation, let $\psi_{i,m}(\Omega_j)$ denote the value taken
 381 by $\psi_{i,m}$ on Ω_j . As \mathcal{B} is orthonormal,

382
$$1 = \langle \psi_{i,m}, \psi_{i,m} \rangle = \sum_{j \in \text{children}(i)} \psi_{i,m}(\Omega_j)^2 v(\Omega_j)$$

383 and so $|\psi_{i,m}(\Omega_j)| \leq v(\Omega_j)^{-1/2}$ for all $j \in \text{children}(i)$. By Theorem 3(a),
 384 $v(\Omega_j)^{-1/2} \leq \underline{B}^{-1/2} v(\Omega_i)^{-1/2}$, and so

385
$$\|\psi_{i,m}\|_{C^0} = \max_{j \in \text{children}(i)} |\psi_{i,m}(\Omega_j)| \leq \underline{B}^{-1/2} v(\Omega_i)^{-1/2}.$$

386 (d) As a result,

387
$$\sum_{m=1}^{\text{nc}(i)-1} \|\psi_{i,m}\|_{C^0} \leq [\text{nc}(i) - 1] \underline{B}^{-1/2} v(\Omega_i)^{-1/2} \leq \underline{B}^{-3/2} v(\Omega_i)^{-1/2}.$$

388 We can now bound $\|f\|_{C^0}$. Take $x_j \in \Omega$. As \mathcal{B} is orthonormal,

389
$$f(x_j) = \hat{f}_{DC} \mathbf{1}_\Omega(x_j) + \sum_{i \in \text{branches}} \sum_{m=1}^{\text{nc}(i)-1} \hat{f}_{i,m} \psi_{i,m}(x_j).$$

390 Recall that $\text{supp}(\psi_{i,m}) \subseteq \Omega_i$, where we define the support of a function to be $\text{supp}(\cdot)$.
 391 By Lemma 1 and Lemma 7(a), $x_j \in \Omega_i$ iff $i \leq j$. So, we can restrict the sum to those
 392 $i \in$ branches with $i \leq j$ (a set not including j , since $j \in$ leaves), obtaining

$$f(x_j) = \hat{f}_{\text{DC}} + \sum_{i < j} \sum_{m=1}^{\text{nc}(i)-1} \hat{f}_{i,m} \psi_{i,m}(x_j).$$

394 Taking absolute values,

$$\begin{aligned} |f(x_j)| &\leq \left| \hat{f}_{\text{DC}} \right| + \sum_{i < j} \max_{1 \leq m < \text{nc}(i)} \left| \hat{f}_{i,m} \right| \sum_{m=1}^{\text{nc}(i)-1} \|\psi_{i,m}\|_{C^0} \\ &\leq \left| \hat{f}_{\text{DC}} \right| + \sum_{i < j} a_i \nu(\Omega_i)^{1/2} \underline{B}^{-3/2} \nu(\Omega_i)^{-1/2} \\ &= \left| \hat{f}_{\text{DC}} \right| + \underline{B}^{-3/2} \sum_{i < j} a_i. \end{aligned}$$

398 Taking the maximum over $x_j \in \Omega$ yields the desired bound. □

399 Consider the coefficient decay condition used in Theorem 3, $\left| \hat{f}_{i,m} \right| \leq a_i \nu(\Omega)^{1/2}$. The
 400 condition cannot be weakened to $\left| \hat{f}_{i,m} \right| \leq C \nu(\Omega_i)^{1/2}$, as shown by Remark 1. On the other
 401 hand, $\left| \hat{f}_{i,m} \right| \leq C \nu(\Omega_i)^{\alpha+1/2}$ is slightly stronger than necessary, as shown by the Theorem 3.
 402 $\left| \hat{f}_{i,m} \right| \leq a_i \nu(\Omega_i)^{1/2}$ lies somewhere in the middle, as expected; all the same, the a_i factor is
 403 somewhat artificial and unsatisfying. For concreteness, we now apply Theorem 3 in the case
 404 that $\left| \hat{f}_{i,m} \right| \leq C \nu(\Omega_i)^{\alpha+1/2}$.

405 **Remark 2** Let $C \geq 0$ and $\alpha \in (0, 1]$. Suppose, as in Theorem 1 and 2, that $\left| \hat{f}_{i,m} \right| \leq$
 406 $C \nu(\Omega_i)^{\alpha+1/2}$ for all wavelets $\psi_{i,m}$. Define $\{a_i : i \in I\}$ by $a_i = C \nu(\Omega_i)^\alpha$, so that $\left| \hat{f}_{i,m} \right| \leq$
 407 $a_i \nu(\Omega_i)^{1/2}$. For any $x_j \in \Omega$,

$$\sum_{i < j} a_i = \sum_{i < j} C \nu(\Omega_i)^\alpha \leq \sum_{i < j} C \underline{B}^{-\alpha \text{depth}(i)} \nu(\Omega)^\alpha \leq \frac{C}{1 - \underline{B}^\alpha}.$$

409 Theorem 3 then gives

$$\|f\|_{C^0} \leq \left| \hat{f}_{\text{DC}} \right| + \max_{x_j \in \Omega} \sum_{i < j} a_i \leq \left| \hat{f}_{\text{DC}} \right| + \frac{C}{\underline{B}^{3/2} (1 - \underline{B}^\alpha)}.$$

411 Unsurprisingly, this bound echoes the bound given in Theorem 1.

412 Theorem 2 used Theorem 1 and Lemma 3 to obtain a thresholding condition for use with
 413 the Hölder norm. Analogously, Theorem 4 uses Theorem 3 to obtain a thresholding condition
 414 for use with the supremum norm.

415 **Theorem 4** Let $\{\Omega_i : i \in I\}$ be a discrete hierarchical tree, and let \mathcal{B} be a Haar-like basis
 416 for V . Let $\{a_i : i \in I\}$ be a collection of nonnegative numbers. Given $f \in V$, define $\tilde{f} \in V$

417 by $\hat{f}_{DC}^\wedge = \hat{f}_{DC}$ and

418
$$\hat{f}_{i,m}^\wedge = \begin{cases} \hat{f}_{i,m} & |\hat{f}_{i,m}| > a_i v(\Omega_i)^{1/2} \\ 0 & \text{otherwise} \end{cases}$$

419 for all wavelets $\psi_{i,m} \in \mathcal{B}$. Then

420
$$\|f - \tilde{f}\|_{C^0} \leq \underline{B}^{-3/2} \max_{x_j \in \Omega} \sum_{i < j} a_i.$$

421 **Proof** By linearity,

422
$$(f - \tilde{f})_{i,m}^\wedge = \begin{cases} 0 & |\hat{f}_{i,m}| > a_i v(\Omega_i)^{1/2} \\ \hat{f}_{i,m} & \text{otherwise} \end{cases}$$

423 for all wavelets $\psi_{i,m} \in \mathcal{B}$. In particular, $|(f - \tilde{f})_{i,m}^\wedge| \leq a_i v(\Omega_i)^{1/2}$. Applying Theorem 3,

424
$$\|f - \tilde{f}\|_{C^0} \leq \underline{B}^{-3/2} \max_{x_j \in \Omega} \sum_{i < j} a_i,$$

425 since $(f - \tilde{f})_{DC}^\wedge = 0$, again by linearity. □

426 We conclude this section with a brief discussion of the smoothness of the thresholding
 427 errors induced in Theorem 2 and 4. In the Hölder norm case, Theorem 2 concludes that
 428 $\|f - \tilde{f}\|_{C^\alpha} \leq \tau$, where τ is independent of the size of the tree. This statement that the
 429 thresholding error is small as measured by the Hölder norm implies that it is smooth as
 430 measured by the Hölder seminorm: as $\|\cdot\|_{C^\alpha} = \|\cdot\|_{C^0} + |\cdot|_{C^\alpha}$, $|f - \tilde{f}|_{C^\alpha} \leq \tau$, i.e.,

431
$$|(f - \tilde{f})(x) - (f - \tilde{f})(y)| \leq \tau d(x, y)^\alpha \quad \text{for all } x, y \in \Omega.$$

432 In the supremum norm case, Theorem 4 concludes that $\|f - \tilde{f}\|_{C^0} \leq \tau$, where τ is
 433 dependent on the size of the tree but can be made independent if, for example, a_i decays
 434 sufficiently quickly with $\text{depth}(i)$. In contrast with the C^α norm, the C^0 norm does not
 435 measure regularity, so this bound implies no smoothness of the thresholding error beyond
 436 the trivial estimate

437
$$|(f - \tilde{f})(x) - (f - \tilde{f})(y)| \leq 2\|f - \tilde{f}\|_{C^0} \leq 2\tau.$$

438 Nonetheless, $f - \tilde{f}$ can be shown to satisfy a natural smoothness condition generalizing
 439 the Hölder condition. Define a *modulus of continuity* to be a function $w : [0, 1] \rightarrow [0, \infty)$
 440 satisfying $\lim_{\epsilon \rightarrow 0} w(\epsilon) = 0$ [39]. We say that a function $f \in V$ admits a modulus of conti-
 441 nuity w if $|f(x) - f(y)| \leq w(d(x, y))$ for all $x, y \in \Omega$. Because Ω is finite, any function
 442 on Ω is automatically uniformly continuous, and so admits some modulus of continuity.
 443 That modulus generally depends on the balance bounds of the tree and the decay rate of the
 444 wavelet coefficients of the function. Lemma 4 details the relationship.

445 **Lemma 4** Let $\{\Omega_i : i \in I\}$ be a discrete hierarchical tree, and let \mathcal{B} be a Haar-like basis for
 446 V . Let $\{a_i : i \in I\}$ be a collection of nonnegative numbers. If $f \in V$ satisfies

447
$$|\hat{f}_{i,m}| \leq a_i v(\Omega_i)^{1/2}$$

448 for all wavelets $\psi_{i,m} \in \mathcal{B}$, then f admits a modulus of continuity depending only on \underline{B} and
 449 the function $s : \mathbb{N} \rightarrow [0, \infty)$ defined by

$$450 \quad s(N) = \max_{x_j \in \Omega} \sum_{\substack{i < j \\ \text{depth}(i) \geq N}} a_i.$$

451 The proof of Lemma 4 is quite similar to the proof of [36, Theorem 2] (Theorem 1 above);
 452 indeed, if $|\hat{f}_{i,m}| \leq C v(\Omega_i)^{\alpha+1/2}$ for some $C \geq 0$ and $\alpha \in (0, 1]$ (as in Theorem 1), the
 453 conclusion of the Lemma 4 (with the dependence on s translated to a dependence on C and
 454 \underline{B}) follows from the Theorem 1 and the definition of the Hölder seminorm.

455 **Proof** Observe that s is nonincreasing, with $s(N) = 0$ for N sufficiently large. Define
 456 $N^* : (0, 1] \rightarrow \mathbb{N}$ by $N^*(\epsilon) = \lceil \log(\epsilon) / \log(\underline{B}) \rceil$. Observe that $N^*(\epsilon) \rightarrow \infty$ as $\epsilon \rightarrow 0$. Define
 457 $w : [0, 1] \rightarrow [0, \infty)$ by $w(0) = 0$ and $w(\epsilon) = 2\underline{B}^{-3/2} s(N^*(\epsilon))$ for $\epsilon \in (0, 1]$. $w(\epsilon) \rightarrow 0$ as
 458 $\epsilon \rightarrow 0$, so w is a modulus of continuity. Observe that w depends only on \underline{B} and s . We claim
 459 that f admits w .

460 Take $x_j, x_{j'} \in \Omega$. If $d(x_j, x_{j'}) = 0$ (i.e., $x_j = x_{j'}$), then automatically $|f(x_j) - f(x_{j'})| =$
 461 $0 = w(\epsilon)$. So, suppose $d(x_j, x_{j'}) \in (0, 1]$. Write $\epsilon = d(x_j, x_{j'})$ and $i^* = \text{lca}(j, j')$, so
 462 that $\epsilon = v(\Omega_{i^*})$. $v(\Omega_{i^*}) \geq \underline{B}^{\text{depth}(i^*)}$, so $\text{depth}(i^*) \geq \log(\epsilon) / \log(\underline{B})$. In particular, since
 463 $\text{depth}(i^*) \in \mathbb{N}$, $\text{depth}(i^*) \geq \lceil \log(\epsilon) / \log(\underline{B}) \rceil = N^*(\epsilon)$. As s is nonincreasing,

$$464 \quad s(\text{depth}(i^*)) \leq s(N^*(\epsilon)). \tag{3}$$

465 For $i_{lo}, i_{hi} \in I$ with $i_{lo} \leq i_{hi}$, let $[i_{lo}, i_{hi})$ denote the interval $\{i \in I : i_{lo} \leq i < i_{hi}\}$. Using
 466 this notation, we can write $f(x_j)$ and $f(x_{j'})$ in terms of the wavelet coefficients as follows:

$$467 \quad f(x_j) = \hat{f}_{DC} + \sum_{i \in [\text{root}, j)} \sum_{m=1}^{\text{nc}(i)-1} \hat{f}_{i,m} \psi_{i,m}(x_j) \quad \text{and}$$

$$468 \quad f(x_{j'}) = \hat{f}_{DC} + \sum_{i \in [\text{root}, j')} \sum_{m=1}^{\text{nc}(i)-1} \hat{f}_{i,m} \psi_{i,m}(x_{j'}).$$

469 Observe that $[\text{root}, j) = [\text{root}, i^*) \sqcup [i^*, j)$ and similarly $[\text{root}, j') = [\text{root}, i^*) \sqcup$
 470 $[i^*, j')$. So, we can write

$$471 \quad f(x_j) - f(x_{j'}) = \underbrace{\sum_{i \in [\text{root}, i^*)} \sum_{m=1}^{\text{nc}(i)-1} \hat{f}_{i,m} [\psi_{i,m}(x_j) - \psi_{i,m}(x_{j'})]}_{(4.i)} \tag{4}$$

$$+ \underbrace{\sum_{i \in [i^*, j)} \sum_{m=1}^{\text{nc}(i)-1} \hat{f}_{i,m} \psi_{i,m}(x_j)}_{(4.ii)} - \underbrace{\sum_{i \in [i^*, j')} \sum_{m=1}^{\text{nc}(i)-1} \hat{f}_{i,m} \psi_{i,m}(x_{j'})}_{(4.iii)}.$$

472 If $i \in [\text{root}, i^*)$, x_j and $x_{j'}$ are contained in the same child of Ω_i (because $i < i^* =$
 473 $\text{lca}(j, j')$), and so $\psi_{i,m}(x_j) = \psi_{i,m}(x_{j'})$ for all $1 \leq m < \text{nc}(i)$. Thus, (4.i) = 0. Next we
 474 bound (4.ii). Taking absolute values,

$$475 \quad |(4.ii)| \leq \sum_{i \in [i^*, j)} \sum_{m=1}^{\text{nc}(i)-1} |\hat{f}_{i,m}| \|\psi_{i,m}\|_{C^0} \leq \sum_{i \in [i^*, j)} \max_{1 \leq m < \text{nc}(i)} |\hat{f}_{i,m}| \sum_{m=1}^{\text{nc}(i)-1} \|\psi_{i,m}\|_{C^0}.$$

476 By hypothesis, each $|\hat{f}_{i,m}|$ is bounded by $a_i v(\Omega_i)^{1/2}$. We found in the proof of Theorem 3
 477 that $\sum_{m=1}^{nc(i)-1} \|\psi_{i,m}\|_{C^0} \leq \underline{B}^{-3/2} v(\Omega_i)^{-1/2}$. As a result,

$$\begin{aligned}
 & |(4.ii)| \leq \sum_{i \in [i^*, j]} a_i v(\Omega_i)^{1/2} \underline{B}^{-3/2} v(\Omega_i)^{-1/2} \\
 & = \underline{B}^{-3/2} \sum_{i \in [i^*, j]} a_i = \underline{B}^{-3/2} \sum_{\substack{i < j \\ \text{depth}(i) \geq \text{depth}(i^*)}} a_i \\
 & \leq \underline{B}^{-3/2} \max_{x_j \in \Omega} \sum_{\substack{i < j \\ \text{depth}(i) \geq \text{depth}(i^*)}} a_i = \underline{B}^{-3/2} s(\text{depth}(i^*)).
 \end{aligned}$$

481 By the same argument, $|(4.iii)| \leq \underline{B}^{-3/2} s(\text{depth}(i^*))$. Combining and applying 3, we arrive
 482 at

$$|f(x_j) - f(x_{j'})| \leq 0 + 2\underline{B}^{-3/2} s(\text{depth}(i^*)) \leq 2\underline{B}^{-3/2} s(N^*(\epsilon)) = w(\epsilon).$$

484 □

485 As an immediate corollary, we can apply Lemma 4 to the thresholding error $f - \tilde{f}$ induced
 486 in Theorem 4. We found in the proof of the Theorem 4 that $|(f - \tilde{f})_{i,m}^\wedge| \leq a_i v(\Omega_i)^{1/2}$. The
 487 Lemma 4 then implies that $f - \tilde{f}$ admits a modulus of continuity depending only on \underline{B} and
 488 $\{a_i : i \in I\}$ (via s).

489 **Remark 3** We noted above that Theorem 2 guarantees that the thresholding error $f - \tilde{f}$
 490 satisfies a smoothness condition independent of the size of the tree in the Hölder norm
 491 case. Does Lemma 4, applied to the thresholding error $f - \tilde{f}$ from Theorem 4, do the
 492 same in the supremum norm case? We consider a specific example for concreteness. Let
 493 $\{\{\Omega_b^n : b \in B^n\} : n \in \mathbb{N}\}$ be the family of hierarchical trees defined in Remark 1. Let
 494 $B = \bigcup_{n \in \mathbb{N}} B^n$, and suppose that a common collection of decay constants $\{a_b : b \in B\}$ is
 495 given. Let a collection of functions $\{f^n : n \in \mathbb{N} \text{ and } f^n \in V^n\}$ be given. Threshold the
 496 functions as in Theorem 4, so that $|(f^n - \tilde{f}^n)_{b,m}^\wedge| \leq a_b v(\Omega_b^n)$. Lemma 4 implies that each
 497 $f^n - \tilde{f}^n$ admits as a modulus of continuity the function w^n defined by

$$w^n(\epsilon) = 2\underline{B}^{-3/2} s^n(N^*(\epsilon)) \quad \text{where} \quad s^n(N) = \max_{x_c \in \Omega^n} \sum_{\substack{b < c \\ \text{depth}(b) \geq N}} a_b.$$

499 We seek a common modulus of continuity w admitted by all of the thresholding errors $f^n - \tilde{f}^n$
 500 simultaneously. The natural candidate is the function $w : [0, 1] \rightarrow [0, \infty)$ defined by

$$w(\epsilon) = 2\underline{B}^{-3/2} \sup_{n \in \mathbb{N}} s^n(N^*(\epsilon)).$$

502 We seek conditions under which (a) w is finite and (b) $\lim_{\epsilon \rightarrow 0} w(\epsilon) = 0$.

503 (a) Since each s^n is decreasing, a sufficient condition is

$$\sup_{n \in \mathbb{N}} s^n(0) = \sup_{n \in \mathbb{N}} \max_{x_c \in \Omega^n} \sum_{b < c} a_b < \infty.$$

505 That is, we require that the sequences $\{a_b : b \in \mathbb{N}, x_c \in \Omega^n, \text{ and } b < c\}$ have uniformly
 506 bounded sums. (Observe that each sequence is automatically summable, since each B^n is

finite.) This condition is not satisfied if, for example, $a_b = 1$ for all $b \in B$. Then $s^n(N) \rightarrow \infty$ as $n \rightarrow \infty$, and so the smoothness property guaranteed by Lemma 4 becomes weaker as the tree grows. Observe that this condition is sufficient for $\|f^n - \tilde{f}^n\|_{C^0}$ to be bounded uniformly, by Theorem 4.

(b) Since $N^*(\epsilon) \rightarrow \infty$ as $\epsilon \rightarrow 0$, a sufficient condition is

$$\lim_{N \rightarrow \infty} \sup_{n \in \mathbb{N}} s^n(N) = \lim_{N \rightarrow \infty} \sup_{n \in \mathbb{N}} \max_{x_c \in \Omega^n} \sum_{\substack{b < c \\ \text{depth}(b) \geq N}} a_b = 0.$$

That is, we require that the partial sums of the sequences $\{a_b : n \in \mathbb{N}, x_c \in \Omega^n, \text{ and } b < c\}$ converge uniformly. This condition is not satisfied if, for example, $a_b = 0$ unless b is of the form $(0, \dots, 0, 1)$, in which case $a_b = 1$. Then (a) above holds (none of the sequences have more than one nonzero term, so all sum to either 0 or 1), but for every $N > 0$ there exists some $n \in \mathbb{N}$ and $x_c \in \Omega^n$ such that $\{a_b : b < c \text{ and } \text{depth}(b) \geq N\}$ sums to 1. For example, take $n = N + 1$ and $c = (0, \dots, 0, 1, 0)$.

4 Numerical Examples

We exercise our method on a set of data compression examples that range from complex boundaries on a piecewise constant function in Sect. 4.1 to classification on the standard MNIST dataset in Sect. 4.3. Theorem 2 and 4 define thresholding algorithms for use with the Hölder and supremum norms. Guaranteed bounds on the magnitude and smoothness of the errors incurred by these techniques were given in the previous Sect. 3. In this section, we complement that theoretical characterization with a trio of numerical examples. In the first, we compute the rate–distortion curve for a piecewise constant function defined on a disk. In the second, we study the gap between the error bound and the achieved error for a specially designed function on an interval. In the third, we generate a family of randomized classifiers on the MNIST dataset and use our thresholding technique to identify those with good generalizability. For the sake of brevity, we restrict our attention to the supremum norm case.

4.1 Piecewise Constant Function

We begin with an investigation of the relationship between the number of wavelet coefficients retained by the thresholding algorithm and the C^0 norm of the error. We take Ω to be a collection of 6708 random points in the unit disk and define a function $f : \Omega \rightarrow \mathbb{R}$ by

$$f(r, \theta) = \begin{cases} 0 & [15r] \equiv 0 \pmod{2} \\ 1 & [15r] \equiv 1 \pmod{2}. \end{cases}$$

In order to apply the thresholding technique described in Theorem 4 to f , we require a discrete hierarchical tree on Ω , a Haar-like basis for V , and a collection of thresholding weights defined on the nodes of the tree. We generate a k -d tree $\{\Omega_i : i \in I\}$ on Ω [40] and verify that it satisfies the definition of a discrete hierarchical tree. We note generation of a k -d tree require a dimension and splitting point for every branch, providing algorithmic freedom in tree generation, but in this paper we cycle through dimensions and use the median point for splitting. Next, we use Algorithm 1 to generate a Haar-like basis \mathcal{B} for V . Finally, we define the thresholding weights $\{a_i : i \in I\}$ by $a_i = Cb^{\text{depth}(i)}$ with $C = 0.05$ and $b = 0.9$.

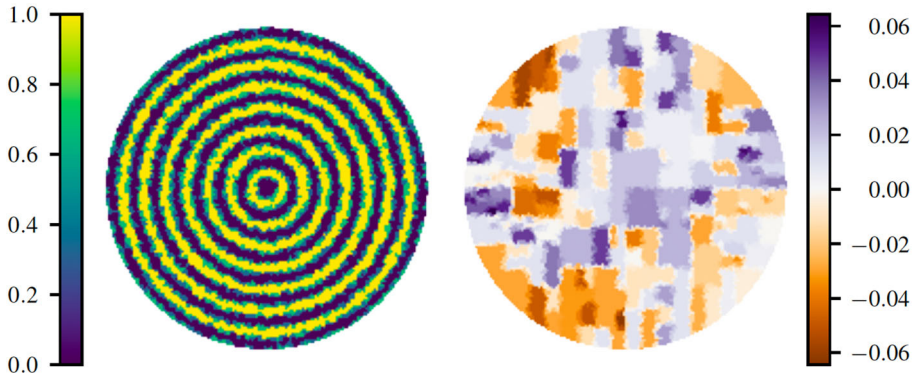


Fig. 3 On the left, the piecewise constant function f defined on Ω , where the green dots represent the location of the points related to the weights retained in thresholding. On the right, the thresholding error $f - \tilde{f}$. \tilde{f} is generated by applying the thresholding technique described in Theorem 4 with thresholding weights $\{a_i : i \in I\}$ defined by $a_i = 0.05 \times 0.9^{\text{depth}(i)}$. The thresholding algorithm retains 3318 of 6707 wavelet coefficients

545 Applying the thresholding technique described in Theorem 4 then yields an approximation
 546 \tilde{f} to f . Of f 's 6707 wavelet coefficients, 3318 are retained to encode \tilde{f} . The supremum
 547 norm of the error is $\|f - \tilde{f}\|_{C^0} \approx 0.0644$. See Fig. 3 for a plot of the input f and the error
 548 $f - \tilde{f}$. The left image depicts the piecewise constant function along with the sampled points
 549 retained by thresholding method, where the right image depicts the thresholding error. It can
 550 be seen that points sampled near the boundaries are retained, which intuitively contain the
 551 most information for reconstructing a piecewise constant function.

552 We next study the effect of changing the thresholding weights on the error. We use the
 553 same domain Ω , function f , hierarchical tree $\{\Omega_i : i \in I\}$, and Haar-like basis \mathcal{B} as before.
 554 We define a parametrized set of thresholding weights $\{a_i : i \in I\}$ by $a_i = Cb^{\text{depth}(i)}$
 555 with $b = 0.9$ unchanged and C variable. For a range of C values, we recompute the thresh-
 556 olding weights, count the fraction of f 's wavelet coefficients that are retained, regenerate
 557 the approximation \tilde{f} , and calculate the achieved error $\|f - \tilde{f}\|_{C^0}$. Plotting, we obtain the
 558 rate–distortion curve shown in Fig. 4. Decreasing C causes more of the wavelet coefficients
 559 to be retained, which generally, but not always (see the caption of Fig. 4), results in lower
 560 achieved error. While the error bound given in Theorem 4 is always respected, it is also rather
 561 pessimistic. We study the gap between the error bound and the achieved error in our next
 562 numerical example.

563 4.2 Thresholding Error Bound Tightness

564 In our second numerical example, we study the tightness of the error bound given in Theorem
 565 4. Take $n \in \mathbb{N}$. Let $\{\Omega_b^n : b \in B^n\}$ be the discrete hierarchical tree and \mathcal{B}^n the Haar-like
 566 basis defined in Remark 1. Define a collection of thresholding weights $\{a_b^n : b \in B^n\}$ by
 567 $a_b^n = 2^{-|b|}$. Let $\text{sgn} : \mathbb{R} \rightarrow \mathbb{R}$ be the sign function, defined by $\text{sgn}(y) = -1$ if $y < 0$,
 568 $\text{sgn}(y) = 0$ if $y = 0$, and $\text{sgn}(y) = 1$ if $y > 0$. Define $f^n : \Omega^n \rightarrow \mathbb{R}$ by $\hat{f}_{\text{DC}}^n = 0$
 569 and $\hat{f}_{b,1}^n = \text{sgn}(\psi_{b,1}^n(0)) a_b^n v(\Omega_b^n)^{1/2}$. Let \tilde{f}^n be the approximation to f^n generated by
 570 applying the thresholding technique described in Theorem 4. For all wavelets $\psi_{b,1}^n \in \mathcal{B}^n$,
 571 $|\hat{f}_{b,1}^n| = a_b^n v(\Omega_b^n)^{1/2} \leq a_b^n v(\Omega_b^n)^{1/2}$, so $\tilde{f}_{b,1}^n = 0$. As additionally $\hat{f}_{\text{DC}}^n = 0$, \tilde{f}^n is

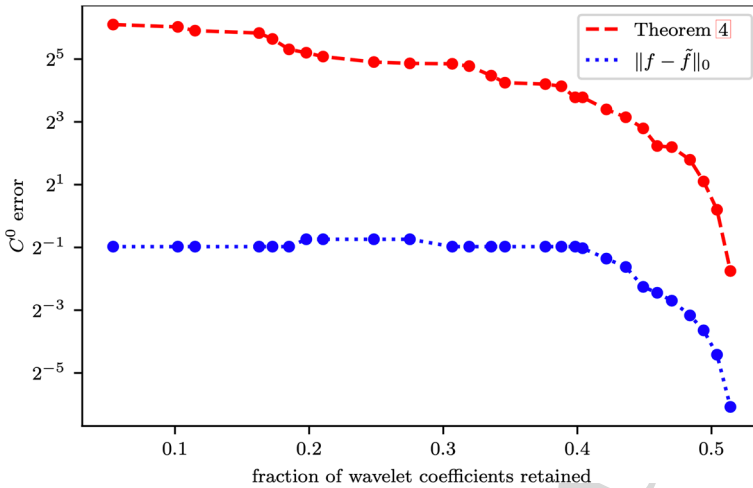


Fig. 4 Rate–distortion curve for the function f plotted in Fig. 3. The error does not decrease monotonically as the number of wavelet coefficients retained increases: it jumps from about 0.599 with 1241 coefficients retained to about 0.599 with 1327 coefficients retained. The error bound is always respected, though it is not especially tight. See Sect. 4.2

the zero function and the thresholding error $f^n - \tilde{f}^n$ is equal to f^n . We next compute the norm $\|f^n - \tilde{f}^n\|_{C^0} = \|f^n\|_{C^0}$ of this error.

Recall the expansion of f^n in terms of its wavelet coefficients: $f^n = \hat{f}_{DC}^n + \sum_{b \in B^{n-1}} \hat{f}_{b,1}^n \psi_{b,1}^n$. The average \hat{f}_{DC}^n is zero. Because of the $\text{sgn}(\psi_{b,1}^n(0))$ factor in its definition, $\hat{f}_{b,1}^n$ is nonzero only if $0 \in \text{supp}(\psi_{b,1}^n) = \Omega_b^n$. As a result,

$$f(x) = \sum_{b < (0, \dots, 0)} \hat{f}_{b,1}^n \psi_{b,1}^n(x) = \sum_{b < (0, \dots, 0)} \text{sgn}(\psi_{b,1}^n(0)) a_b^n \nu(\Omega_b^n) \psi_{b,1}^n(x).$$

The magnitude of the sum is maximized when $\psi_{b,1}^n(0)$ and $\psi_{b,1}^n(x)$ have the same sign. In particular, $\|f^n - \tilde{f}^n\|_{C^0} = \|f^n\|_{C^0} = f(0)$. $a_b^n = 2^{-|b|}$, $\nu(\Omega_b^n)^{1/2} = 2^{-|b|/2}$, and $\psi_{b,1}^n(0) = -2^{|b|/2}$ if $0 \in \Omega_b^n$, so

$$\|f^n - \tilde{f}^n\|_{C^0} = f(0) = \sum_{b < (0, \dots, 0)} 2^{-|b|} 2^{-|b|/2} 2^{|b|/2} = \sum_{k=0}^{n-1} 2^{-k} = 2(1 - 2^{-n}).$$

The error bound given in Theorem 4, meanwhile, is

$$\|f^n - \tilde{f}^n\|_{C^0} \leq B^{-3/2} \max_{x_c \in \Omega^n} \sum_{b < c} a_b^n = 2^{3/2} \sum_{k=0}^{n-1} 2^{-k} = 4\sqrt{2}(1 - 2^{-n}).$$

f^n , the achieved error, and the error bound are plotted in Fig. 5 in the case $n = 10$. The remainder of this section is an account of the factor $2\sqrt{2}$ separating the error bound from the achieved error.

Remark 4 The gap between the error bound and achieved error in the previous numerical example is due to two bounds in the proof of Theorem 3 that are loose for $\{\Omega_b^n : b \in B^n\}$.

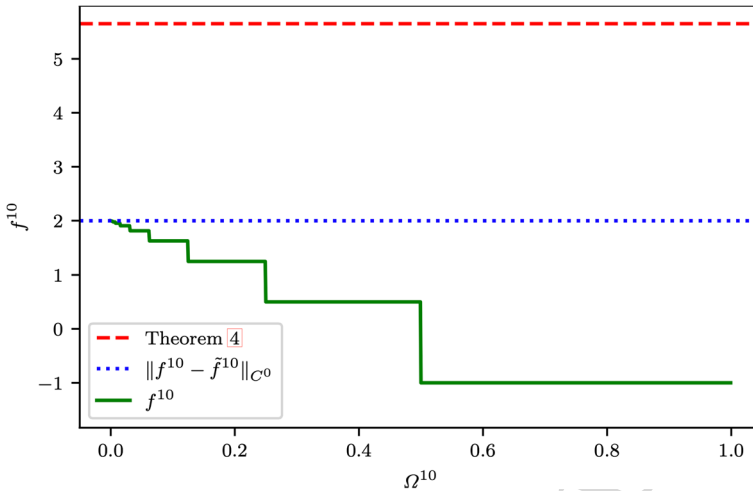


Fig. 5 Illustration of the gap between the error bound given in Theorem 4 and the achieved error in the case of the function f^{10} defined on Ω^{10} . f^{10} attains its maximum magnitude at 0. The supremum norm of the error is $\|f^{10} - \tilde{f}^{10}\|_{C^0} = 2(1 - 2^{-10})$. The error bound given in Theorem 4 is $\underline{B}^{-3/2} \max_{x_c \in \Omega^{10}} \sum_{b < c} a_b^{10} = 4\sqrt{2}(1 - 2^{-10})$

589 We summarize those bounds here for ease of reference. $\{\Omega_i : i \in I\}$ is a discrete hierarchical
 590 tree, \mathcal{B} is a Haar-like basis for V , and $i \in \text{branches}$.

591 (b) Writing $\text{nc}(i) = |\text{children}(i)|$, $\text{nc}(i) - 1 \leq \text{nc}(i) \leq \underline{B}^{-1}$.

592 (c) Let $\psi_{i,m}$ be any of the wavelets $\psi_{i,1}, \dots, \psi_{i,\text{nc}(i)-1}$. Then

593
$$\|\psi_{i,m}\|_{C^0} \leq \underline{B}^{-1/2} v(\Omega_i)^{-1/2}.$$

594 If $\{\Omega_i : i \in I\}$ is taken to be $\{\Omega_b^n : b \in B^n\}$, then $\text{nc}(i) = 2$ and $\underline{B} = \frac{1}{2}$. Theorem 3(b)
 595 then bounds $\text{nc}(i) - 1 = 1$ by $\underline{B}^{-1} = 2$. The source of the looseness is the first inequality,
 596 $\text{nc}(i) - 1 \leq \text{nc}(i)$, which is obviously never tight. $\text{nc}(i) = 2$ is the worst case scenario
 597 ($\text{nc}(i) = 1$ is disallowed by Definition 1 (b)), and the ratio of $\text{nc}(i) - 1$ to $\text{nc}(i)$ approaches
 598 1 as the number of children grows. Consider next the second inequality, $\text{nc}(i) \leq \underline{B}^{-1}$.
 599 We show in Appendix C (see Lemma 12) that $\text{nc}(i) = \underline{B}^{-1}$ if $v(\Omega_j) = \underline{B}v(\Omega_i)$ for all
 600 $j \in \text{children}(i)$. So, this inequality is tight if $\{\Omega_i : i \in I\}$ is a balanced, full k -ary tree.
 601 We may therefore say that Theorem 3(b) is asymptotically tight in the case of balanced, full
 602 k -ary trees as $k \rightarrow \infty$.

603 If $\{\Omega_i : i \in I\}$ is taken to be $\{\Omega_b^n : b \in B^n\}$, then $\underline{B} = \frac{1}{2}$ and $v(\Omega_i) = 2^{-\text{depth}(i)}$. If in
 604 addition \mathcal{B} is taken to be \mathcal{B}^n , then $\psi_{i,1} = -2^{\text{depth}(i)/2} \mathbf{1}_{\Omega_1} + 2^{\text{depth}(i)/2} \mathbf{1}_{\Omega_2}$ with Ω_1 and Ω_2
 605 the appropriately ordered children of Ω_i . Theorem 3(c) then bounds $\|\psi_{i,1}\|_{C^0} = 2^{\text{depth}(i)/2}$
 606 by $\underline{B}^{-1/2} v(\Omega_i)^{-1/2} = 2^{\text{depth}(i)/2} \sqrt{2}$. The source of the looseness is a failure to use the
 607 property that $\psi_{i,1}$ has average zero. We use this property to obtain an improved bound in
 608 Appendix C (see Lemma 13). In the case of a balanced, full k -ary tree $\{\Omega_i : i \in I\}$, Eq. 5,
 609 which is tight, reads

610
$$\|\psi_{i,m}\|_{C^0} \leq \sqrt{\left[\min_{j \in \text{children}(i)} v(\Omega_j) \right]^{-1} - \left[v(\Omega_i) \right]^{-1}} = \sqrt{k-1} v(\Omega_i)^{-1/2}.$$

For a balanced, full k -ary tree, $B^{-1/2} = \sqrt{k}$. so we may say that Theorem 3(c), like 3(b), is asymptotically tight in the case of balanced, full k -ary trees as $k \rightarrow \infty$.

In conclusion, the gap between the error bound and the achieve error seen in Fig. 5 is entirely explained by two suboptimal bounds in the proof of Theorem 3. These bounds are loose for the function f^{10} depicted in Fig. 5, but they are asymptotically tight in the case of balanced, full k -ary trees as $k \rightarrow \infty$.

4.3 MNIST Database

In our third numerical example, we use our thresholding technique to identify classifiers with good generalizability on the MNIST handwritten digit database [41]. The experiment is structured as follows. The data are first preprocessed, and then the training set is used to define 1000 randomized classifiers. For each of these classifiers, a family of reduced approximations is generated with the Theorem 4 thresholding technique. Finally, the classification accuracies achieved by the families on each of the training and testing sets are measured and are found to be strongly correlated.

We begin by reducing the dimensionality of the data. The database comprises 50 000 training and 10 000 testing images of the digits 0–9. All images in the database consist of images of size 28×28 pixels. Following Lepelaars [42], we compute a truncated SVD of the training set, into a matrix of size $28^2 \times 50000$, and project the data onto the span of the right singular vectors corresponding to the 50 largest singular values [43, 44]. We then embed these projections into a space of dimension 50 using t -distributed stochastic neighbor embedding (t -SNE) [45, 46]. In the embedding space, the instances are effectively clustered according to their label. See [42] for a representative illustration made using a slightly different dimensionality reduction procedure. Thus our dataset for classification consists of 50 largest singular values mapped to the digits 0–9.

Next, we build a forest of 1000 randomized k -d trees on the embedded feature vectors using a slightly modified version of Algorithm 2. As written, the Algorithm 2 is not guaranteed to produce a hierarchical tree, as $\Omega_{b+(0)}$ in Algorithm 2 Line 15 will be empty if the median of $\{x_i : x \in \Omega_b\}$ is equal to that set's minimum. The Algorithm 2 used in our implementation slightly shifts the dividing hyperplane in this case. Each k -d tree defines a classifier as follows. The tree defines a partition of the embedding space, with each leaf in the tree associated to a set in the partition. An image to be classified is located in one of these sets, and the training image contained in the corresponding leaf (see Lemma 10) is identified. The label of this training instance is the output classification. By construction, these classifiers achieve 100 % accuracy on the training set. The accuracies achieved on the testing set vary.

After the classifiers are defined, they are thresholded using the technique described in Theorem 4. A Haar-like basis is generated from each tree using Algorithm 1, and the corresponding classifier is decomposed with respect to this basis. We define thresholding weights $\{a_i : i \in I\}$ by $a_i = Cb^{\text{depth}(i)}$ with $b = 0.9$ and C chosen so that 1 %, 2 %, . . . , 19 %, 20 %, 25 %, . . . , 90 %, 95 % of the basis coefficients are retained.³ The thresholded classifiers are then used to classify the images in the training and testing sets. Figure 6 shows the distribution of the accuracies obtained when 11 % of the coefficients are retained. We find that the accuracy of a thresholded classifier on the training set is highly correlated to its accuracy on the testing set.

³ We sample [20 %, 100 %] with lower resolution because we observe very little change in the behavior of the classifiers between the 20 % and 100 % (unthresholded) coefficient retention levels.

Algorithm 2 Generation of a randomized k-d tree from a set of feature vectors. We reuse the concatenation notation $b + b'$ from Remark 1. Ω is assumed to be nonempty and finite.

```

1: function GENERATERANDOMIZEDKDTREE(set  $\Omega$  of feature vectors in  $\mathbb{R}^k$ )
2:   initialize  $B$  to {}
3:   initialize splitting_queue to {}
4:   denoting by root the empty sequence (), define  $\Omega_{\text{root}}$  to be  $\Omega$  and add root
5:     to splitting_queue
6:   while |splitting_queue| > 0 do
7:     pop an element  $b$  from splitting_queue and add it to  $B$ 
8:     if  $|\Omega_b| = 1$  then
9:       continue
10:    end if
11:    randomly choose with uniform probability an element
12:       $i$  from  $\{i : 1 \leq i \leq k \text{ and } |\{x_i : x \in \Omega_b\}| > 1\}$ 
13:    find the median  $m$  of  $\{x_i : x \in \Omega_b\}$ 
14:    define  $\Omega_{b+(0)}$  to be  $\{x : x \in \Omega_b \text{ and } x_i < m\}$  and add  $b + (0)$  to
15:      splitting_queue
16:    define  $\Omega_{b+(1)}$  to be  $\{x : x \in \Omega_b \text{ and } x_i \geq m\}$  and add  $b + (1)$  to
17:      splitting_queue
18:  end while
19:  return  $\{\Omega_b : b \in B\}$ 
20: end function

```

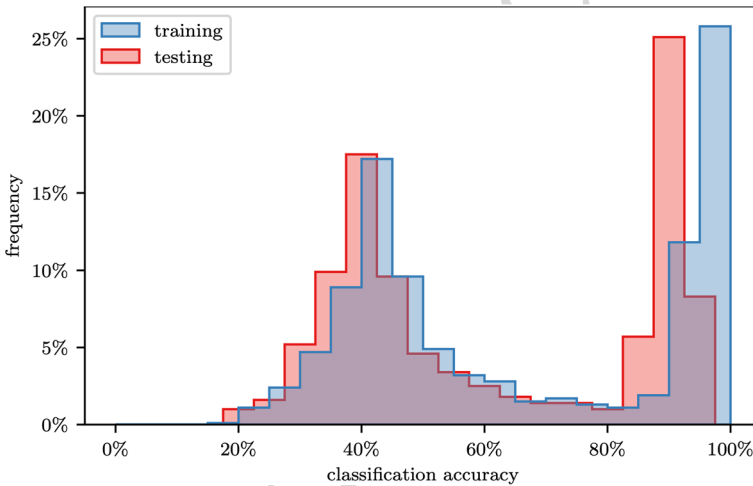


Fig. 6 Illustration of the classification accuracies achieved by 1000 thresholded classifiers on the MNIST training and testing sets. Each classifier is thresholded so that 11 % of its coefficients are retained. The testing set accuracies are slightly lower than the training set accuracies, but otherwise the distributions are similar. In fact, the two accuracy measures are highly correlated: the Pearson correlation coefficient between them is $r \approx 0.9996$

654 Finally, we investigate whether this relationship is particular to the 11 % coefficient retention
655 level or whether it holds more generally. To do this, we introduce the *integrated accuracy*,
656 a simple measure of accuracy across a range of retention levels. To compute the integrated
657 accuracy of a classifier on a dataset, we measure its accuracy on that set before thresholding
658 (at the 100 % level) and also at each of the 35 levels given above (from 1 % to 95 %). We then
659 integrate these accuracies using the trapezoidal rule, yielding the integrated accuracy. We
660 compute the integrated accuracies of all 1000 classifiers on the training set, then normalize

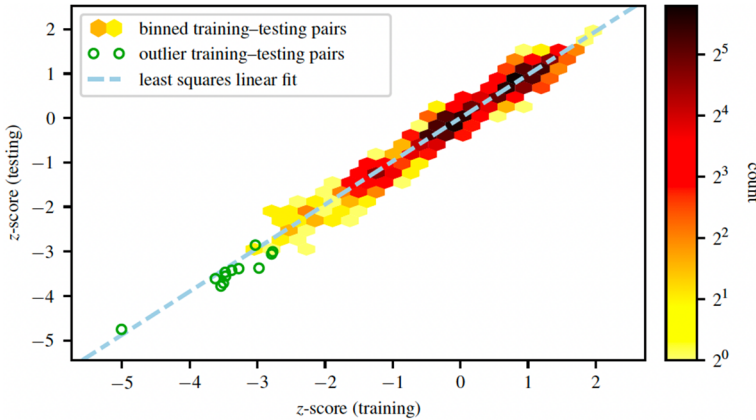


Fig. 7 Illustration of the correlation between the classifiers’ integrated accuracies on the MNIST training and testing sets. The color of each hexagon indicates the number of training–testing pairs contained in it. A least squares linear regression is performed on the pairs, with outliers excluded. The coefficient of determination of the linear model is $r^2 \approx 0.95773$

661 by subtracting the sample mean and dividing by the sample standard deviation. We do the
 662 same on the testing set and then perform a least squares linear regression between the two
 663 measures, excluding outliers (training–testing pairs more than three standard deviations from
 664 the mean in either dimension). The fit is shown in Fig. 7. The coefficient of determination of
 665 the regression is high ($r^2 \approx 0.95773$), indicating that a classifier’s integrated accuracy on the
 666 training set is highly predictive of its integrated accuracy on the testing set. This suggests that
 667 classifiers that perform relatively well on the training set after being thresholded are likely
 668 to generalize relatively well to novel instances.

669 **Funding** This material is based upon work supported by the US Department of Energy, Office of Science, Office
 670 of Advanced Scientific Computing Research, Scientific Discovery through Advanced Computing (SciDAC)
 671 program under the FASTMath institute and the scientific data compression project.

672 **Data Availability** The MNIST dataset used in Sect. 4.3 is available at Yann LeCun’s website (<http://yann.lecun.com/exdb/mnist/>). The datasets used in Sect. 4.1 and 4.2 are available from the corresponding author
 673 upon request.
 674

675 **Declarations**

676 **Conflict of interest** The authors have no competing interests to declare that are relevant to the content of this
 677 article.

678 **A Hierarchical Trees**

679 **Lemma 5** Let (N, E, n_{root}) be a rooted tree. Define a relation \preceq on N by $n \preceq m$ iff the path
 680 from n_{root} to m goes through n . \preceq is a partial order on N .

681 **Proof** We must show that, for $n, m, o \in N$, (a) $n \preceq n$, (b) $n = m$ if $n \preceq m$ and $n \succeq m$, and
 682 (c) $n \preceq o$ if $n \preceq m$ and $m \preceq o$.

683 (a) The path from n_{root} to n necessarily includes n , so $n \preceq n$.

- 684 (b) Suppose $n \neq m$. The path from n_{root} to m includes n , since $n \leq m$. In particular, it
 685 contains as a subset a path from n_{root} to n not including m . On the other hand, the path
 686 from n_{root} to n includes m , since $m \leq n$. There are therefore two distinct paths from
 687 n_{root} to n , one including m and one not. This contradicts the definition of a tree.
 688 (c) The path from n_{root} to o includes m , since $m \leq o$. This path contains a path from n_{root}
 689 to m , which must include n , since $n \leq m$. So, the path from n_{root} to o also includes n ,
 690 and so $n \leq o$.

691 □

692 **Lemma 6** Let $\{\Omega_i : i \in I\}$ be a hierarchical tree. For $j, j' \in I$, $\text{lca}(j, j')$ is well-defined
 693 by Definition 3.

694 **Proof** Let C be the collection of indices $\{i \in I : i \leq j, j'\}$. We claim that C has a unique
 695 element of maximal depth. First, note that C can have only one element at a particular depth,
 696 because j (or, equally well, j') can have only one ancestor at a particular depth. So, it suffices
 697 to show the existence of some element of maximal depth.

698 $\{\text{depth}(i) : i \in C\}$ is bounded: if $i \in C$, then $\text{depth}(i) \leq \text{depth}(j), \text{depth}(j')$,
 699 since $i \leq j, j'$. Furthermore, C is nonempty, because $\text{root} \leq j, j'$ by definition. There
 700 therefore exists some element of maximal depth. □

701 Given $i, i' \in I$, write $i \parallel i'$ if $i \not\leq i'$ and $i' \not\leq i$.

702 **Lemma 7** Let $\{\Omega_i : i \in I\}$ be a hierarchical tree. Let $i, i' \in I$.

703 (a) $i \leq i'$ iff $\Omega_i \supseteq \Omega_{i'}$.

704 (b) $i \parallel i'$ iff $\Omega_i \cap \Omega_{i'} = \emptyset$.

705 **Proof** We begin with a useful intermediate result. Assume the forward direction of (a).
 706 Let $i^* = \text{lca}(i, i')$. We claim that if $i^* \neq i, i'$, then $\Omega_i \cap \Omega_{i'} = \emptyset$. Let $n =$
 707 $\text{depth}(i) - \text{depth}(i^*)$ and $n' = \text{depth}(i') - \text{depth}(i^*)$. As $i^* < i, i', n, n' \geq 1$.
 708 $\text{parent}^{n-1}(\Omega_i)$ and $\text{parent}^{n'-1}(\Omega_{i'})$ are then well-defined. These sets are children of
 709 $\Omega_{i^*} = \text{parent}^n(\Omega_i) = \text{parent}^{n'}(\Omega_{i'})$. By the definition of the lowest common ancestor,
 710 they must be distinct. By Definition 1 (b), then, they must be disjoint. $\text{parent}^{n-1}(\Omega_i) \supseteq \Omega_i$
 711 and $\text{parent}^{n'-1}(\Omega_{i'}) \supseteq \Omega_{i'}$ by the forward direction of (a). So, $\Omega_i \cap \Omega_{i'}$ is a subset of
 712 $\text{parent}^{n-1}(\Omega_i) \cap \text{parent}^{n'-1}(\Omega_{i'})$. The latter is empty, so the former must be empty.

713 (a) Note that the forward direction does not depend on the intermediate result, which assumes
 714 it.

715 (\implies) If $i \leq i'$, then $i = \text{parent}^n(i')$ with $n = \text{depth}(i') - \text{depth}(i)$. Definition
 716 1 (b) dictates that each parent contain its children. Applying repeatedly, we have

$$717 \Omega_i = \text{parent}^n(\Omega_{i'}) \supseteq \dots \supseteq \text{parent}^1(\Omega_{i'}) \supseteq \Omega_{i'}.$$

718
 719 (\impliedby) Let $i^* = \text{lca}(i, i')$.
 720 $i^* \leq i'$, so we are done if $i = i^*$.

721 So, suppose $i \neq i^*$.
 722 Suppose $i' = i^*$. $\Omega_{i^*} \supseteq \Omega_i$ by the forward direction. $\Omega_i \supseteq \Omega_{i'} = \Omega_{i^*}$ by assumption, so
 723 we have $\Omega_i = \Omega_{i^*}$. This is a violation of Definition 1 (b), which dictates that descendants
 724 be strict subsets of their ancestors.

725 So, suppose $i' \neq i^*$. $i^* \neq i, i'$, so by the intermediate result $\Omega_i \cap \Omega_{i'} = \Omega_{i'} = \emptyset$. By
 726 the definition of a hierarchical tree, though, $\Omega_{i'}$ must be nonempty.

727 (b) (\implies) Let $i^* = \text{lca}(i, i')$. Since $i \parallel i'$, $i^* \neq i, i'$. The conclusion then follows from
 728 the intermediate result.
 729 (\impliedby) Since $\Omega_{i'}$ is nonempty by the definition of a hierarchical tree, $\Omega_i \not\subseteq \Omega_{i'}$, and so
 730 $i \not\leq i'$ by the forward direction of (a). By the same argument, $i \not\geq i'$. □

731
 732 **Corollary 1** Let $\{\Omega_i : i \in I\}$ be a hierarchical tree. Let $i, i' \in I$. If $\Omega_i \cap \Omega_{i'} \neq \emptyset$ and
 733 $\Omega_i \not\subseteq \Omega_{i'}$, then $i \leq i'$.

734 **Proof** $\Omega_i \cap \Omega_{i'} \neq \emptyset$ implies $i \leq i'$ or $i \geq i'$ by Lemma 7(b). $\Omega_i \not\subseteq \Omega_{i'}$ implies $i \not\leq i'$ by
 735 Lemma 7(a). Therefore $i \leq i'$. □

736 **Lemma 8** Let $\{\Omega_i : i \in I\}$ be a hierarchical tree. For $x, y \in \Omega$ distinct, $\text{lca}(x, y)$ is
 737 well-defined by Definition 3.

738 **Proof** Let C be the collection of sets $\{\Omega_i : i \in I \text{ and } x, y \in \Omega_i\}$. We claim that C has a
 739 unique element of maximal depth.

740 We begin with existence. First, note that C is nonempty: $\Omega_{\text{root}} \in C$ by Definition 1 (a). By
 741 Definition 1 (c), there exists some $i' \in I$ such that $\Omega_{i'} \ni x$ but $\Omega_{i'} \not\ni y$. Let Ω_i be an element
 742 of C . $\Omega_i \cap \Omega_{i'} \neq \emptyset$, since both Ω_i and $\Omega_{i'}$ contain x . On the other hand, since $\Omega_{i'}$ does not
 743 contain y , $\Omega_i \not\subseteq \Omega_{i'}$. By Corollary 1, then, $i \leq i'$. In particular, $\text{depth}(i) \leq \text{depth}(i')$,
 744 and so $\{\text{depth}(\Omega_i) : \Omega_i \in C\}$ is bounded. As a result, there exists at least one element of
 745 maximal depth.

746 To show uniqueness, suppose there exist two elements of maximal depth, Ω_i and $\Omega_{i'}$.
 747 $\Omega_i \cap \Omega_{i'} \neq \emptyset$, since both Ω_i and $\Omega_{i'}$ contain x (and y). By Lemma 7(b), then, $i \leq i'$ or
 748 $i' \leq i$. In either case, since $\text{depth}(i) = \text{depth}(i')$, $i = i'$. The element of maximal depth
 749 is therefore unique. □

750 **Corollary 2** Let $\{\Omega_i : i \in I\}$ be a hierarchical tree. Let $x, y \in \Omega$ be distinct, and let $i \in I$.
 751 If $\Omega_i \ni x, y$, then $\Omega_i \leq \text{lca}(x, y)$.

752 **Proof** Write $\Omega_{i^*} = \text{lca}(x, y)$. By definition, $\Omega_{i^*} \ni x, y$. As $\Omega_i \ni x, y$ by assumption, Ω_i
 753 and Ω_{i^*} intersect, and so $i \leq i^*$ or $i^* \leq i$ by Lemma 7(b). If $i^* < i$, then $\text{depth}(i^*) <$
 754 $\text{depth}(i)$, contradicting the definition of the lowest common ancestor. Therefore $i \leq i^*$. □

755 **Lemma 9** Let $\{\Omega_i : i \in I\}$ be a hierarchical tree. d is an ultrametric on Ω . i.e., for $x, y, z \in$
 756 Ω , then $d(x, z) \leq \max\{d(x, y), d(y, z)\}$.

757 One approach to proving Lemma 9 is to realize the hierarchical tree as an *ultrametric*
 758 *tree* [47] where the edge between Ω_i and its child Ω_j has weight $[v(\Omega_i) - v(\Omega_j)]/2$ if
 759 $j \in \text{branches}$ and $v(\Omega_i)/2$ if $j \in \text{leaves}$. d is then the metric induced by the edge
 760 weights. A proof which relies instead on the structure of hierarchical trees follows.

761 **Proof** Let $x, y, z \in \Omega$. We must show that (a) d is nonnegative, (b) d is symmetric, (c)
 762 $d(x, y) = 0$ iff $x = y$, and (d) the ultrametric inequality $d(x, z) \leq \max\{d(x, y), d(y, z)\}$
 763 holds.

- 764 (a) Nonnegativity follows from the nonnegativity of v .
- 765 (b) Symmetry follows from the symmetry of lca .
- 766 (c) If $x = y$, then $d(x, y) = 0$ by definition. If $x \neq y$, then $d(x, y) = v(\text{lca}(x, y))$. In
 767 the discrete case, $\text{lca}(x, y)$ is nonempty and v is a rescaling of the counting measure,
 768 so $v(\text{lca}(x, y)) \neq 0$. In the continuous case, $\text{lca}(x, y)$ has nonzero Lebesgue measure
 769 and v is a rescaling of the Lebesgue measure, so again $v(\text{lca}(x, y)) \neq 0$.

(d) If the points are not distinct or if $d(x, z) \leq d(x, y)$, then the inequality holds automatically. So, suppose the points are distinct and $d(x, z) > d(x, y)$. Write $\Omega_{i_{xy}^*} = \text{lca}(x, y)$, $\Omega_{i_{xz}^*} = \text{lca}(x, z)$, and $\Omega_{i_{yz}^*} = \text{lca}(y, z)$. We aim to show that $d(x, z)$ is equal to $d(y, z)$. As $d(x, z) = v(\Omega_{i_{xz}^*})$ and $d(y, z) = v(\Omega_{i_{yz}^*})$, it suffices to show that $i_{xz}^* = i_{yz}^*$. We claim that $i_{xz}^* \leq i_{xy}^*$. $\Omega_{i_{xz}^*}$ and $\Omega_{i_{xy}^*}$ intersect, since both contain x . Because v is a measure, $v(\Omega_{i_{xz}^*}) \leq v(\Omega_{i_{xy}^*})$ if $\Omega_{i_{xz}^*} \subseteq \Omega_{i_{xy}^*}$. By assumption, though, $d(x, z) = v(\Omega_{i_{xz}^*}) > v(\Omega_{i_{xy}^*}) = d(x, y)$. Therefore $\Omega_{i_{xz}^*} \not\subseteq \Omega_{i_{xy}^*}$. By Corollary 1, then, $i_{xz}^* \leq i_{xy}^*$. In particular, since $i_{xz}^* \neq i_{xy}^*$, $i_{xz}^* < i_{xy}^*$. We claim that $i_{xz}^* \leq i_{yz}^*$. $\Omega_{i_{xz}^*} \ni z$ and $\Omega_{i_{yz}^*} \ni y$ automatically. Using Lemma 7(a), since $i_{xz}^* \leq i_{xy}^*$, $\Omega_{i_{xz}^*} \supseteq \Omega_{i_{yz}^*}$. As a result, $\Omega_{i_{xz}^*} \ni y, z$, and so $i_{xz}^* \leq i_{yz}^*$ by Corollary 2. We claim that $\Omega_{i_{xy}^*} \not\ni z$. $\Omega_{i_{xy}^*} \ni x$; if in addition $\Omega_{i_{xy}^*} \ni z$, then $i_{xy}^* \leq i_{xz}^*$ by Corollary 2. But we know that $i_{xz}^* < i_{xy}^*$, so in fact $\Omega_{i_{xy}^*} \not\ni z$. We claim that $i_{yz}^* \leq i_{xy}^*$. $\Omega_{i_{yz}^*}$ and $\Omega_{i_{xy}^*}$ intersect, since both contain y . $\Omega_{i_{yz}^*} \ni z$, but $\Omega_{i_{xy}^*} \not\ni z$, as shown in the previous paragraph. That is, $\Omega_{i_{yz}^*} \not\subseteq \Omega_{i_{xy}^*}$. By Corollary 1, then, $i_{yz}^* \leq i_{xy}^*$. We claim that $i_{yz}^* \leq i_{xz}^*$. $\Omega_{i_{yz}^*} \ni y$ automatically. Because $i_{yz}^* \leq i_{xy}^*$ and $\Omega_{i_{xy}^*} \ni x$, $\Omega_{i_{yz}^*} \ni x$. By Corollary 2, then, $i_{yz}^* \leq i_{xz}^*$. We conclude that $i_{xz}^* = i_{yz}^*$, so that the ultrametric inequality holds.

□

B Discrete Hierarchical Trees

Lemma 10 Let $\{\Omega_i : i \in I\}$ be a discrete hierarchical tree. $|\Omega_i| = 1$ for all $i \in \text{leaves}$.

Proof Let $i \in \text{leaves}$, and suppose $|\Omega_i| \neq 1$. Ω_i is nonempty, so $|\Omega_i| > 1$. In particular, there exist distinct $x, y \in \Omega_i$. By Definition 1 (c), there exists some $i' \in I$ with $x \in \Omega_{i'}$ and $y \notin \Omega_{i'}$. Ω_i and $\Omega_{i'}$ have nonempty intersection, since both contain x . On the other hand, $\Omega_{i'}$ is not a superset of Ω_i , since the latter contains y and the former does not. So, by Corollary 1, $i \leq i'$. In particular, $\text{children}(i)$ is not empty. This contradicts the inclusion of i in leaves. □

Lemma 11 Let $\{\Omega_i : i \in I\}$ be a discrete hierarchical tree. For all $x \in \Omega$, there exists some $i \in \text{leaves}$ such that $\Omega_i \ni x$.

Proof Suppose there exists some $x \in \Omega$ such that there exists no $i \in \text{leaves}$ with $\Omega_i \ni x$. If $i \in I$ satisfies $\Omega_i \ni x$, then, $i \notin \text{leaves}$, and so $\text{children}(i)$ is nonempty. In particular, there exists $j \in \text{children}(i)$ with $\Omega_j \ni x$, by Definition 1 (b). Observe that $\text{depth}(\Omega_j) = \text{depth}(\Omega_i) + 1$. That is, for all $i \in I$ with $\Omega_i \ni x$, there exists some $j \in I$ with $\Omega_j \ni x$ and $\text{depth}(\Omega_j) = \text{depth}(\Omega_i) + 1$. Furthermore, there does exist at least one $i \in I$ (namely, root) with $\Omega_i \ni x$, by Definition 1 (a). The set $\{\text{depth}(\Omega_i) : i \in I \text{ and } \Omega_i \ni x\}$ is therefore unbounded.

We now show that $\{\text{depth}(\Omega_i) : i \in I\}$ is in fact bounded, so that no such $x \in \Omega$ exists. We claim that $\text{depth}(\Omega_i) \leq |\Omega| - |\Omega_i|$ for all $i \in I$. If $i = \text{root}$, then

$$\text{depth}(\Omega_i) = 0 = |\Omega| - |\Omega_i|$$

since $\Omega_{\text{root}} = \Omega$ by Definition 1 (a). Otherwise, observe that $|\Omega_i| \leq |\text{parent}(\Omega_i)| - 1$ by Definition 1 (b). Applying this inequality repeatedly, we have

$$\text{depth}(\Omega_i) \leq |\text{parent}(\Omega_i)| - 1 \leq \dots \leq |\text{parent}^n(\Omega_i)| - n$$

for $n \leq \text{depth}(\Omega_i)$. Setting $n = \text{depth}(\Omega_i)$, we obtain $|\Omega_i| \leq |\Omega| - \text{depth}(\Omega_i)$ (i.e., $\text{depth}(\Omega_i) \leq |\Omega| - |\Omega_i|$), since then $\text{parent}^n(\Omega_i) = \Omega_{\text{root}} = \Omega$. As a result, $\{\text{depth}(\Omega_i) : i \in I\}$ is bounded, and so there exists some $i \in \text{leaves}$ with $\Omega_i \ni x$ for any $x \in \Omega$. □

C Lemmas for Remark 4

Lemma 12 *Let $\{\Omega_i : i \in I\}$ be a discrete hierarchical tree. Let $i \in \text{branches}$. $\text{nc}(i) = \underline{B}^{-1}$ iff $v(\Omega_j) = \underline{B}v(\Omega_i)$ for all $j \in \text{children}(i)$.*

Proof

(\Leftarrow) Because v is a measure and Ω_i is the disjoint union of its children,

$$v(\Omega_i) = \sum_{j=1}^{\text{nc}(i)} v(\Omega_j) = \sum_{j=1}^{\text{nc}(i)} \underline{B}v(\Omega_i) = \underline{B}\text{nc}(i)v(\Omega_i).$$

Therefore, $\text{nc}(i) = \underline{B}^{-1}$.

(\Rightarrow) By the definition of \underline{B} , $v(\Omega_j) \geq \underline{B}v(\Omega_i)$ for all $j \in \text{children}(i)$. The reverse inequality also holds: taking Ω_1 as an example,

$$\begin{aligned} v(\Omega_1) &= v(\Omega_i) - \sum_{j=2}^{\text{nc}(i)} v(\Omega_j) \\ &\leq v(\Omega_i) - \underline{B}[\text{nc}(i) - 1]v(\Omega_i) \\ &= \underline{B}v(\Omega_i). \end{aligned}$$

since $\text{nc}(i) = \underline{B}^{-1}$. As a result, $v(\Omega_1) = \underline{B}v(\Omega_i)$, and the same holds for the other children of Ω_i . □

Lemma 13 *Let $\{\Omega_i : i \in I\}$ be a discrete hierarchical tree. If $\psi_{i,m}$ is a wavelet of a Haar-like basis for V , then*

$$\|\psi_{i,m}\|_{C^0} \leq \sqrt{\frac{1}{\min_{j \in \text{children}(i)} v(\Omega_j)} - \frac{1}{v(\Omega_i)}}. \tag{5}$$

Furthermore, this bound is tight.

Proof For $i \in \text{branches}$, denote by V_i the linear span of $\{\mathbf{1}_{\Omega_j} : j \in \text{children}(i)\}$ and by W_i the space $V_i \cap \mathbf{1}_{\Omega_i}^\perp$. By Definition 2, if $\psi_{i,m}$ is a wavelet of a Haar-like basis, then $\psi_{i,m}$ has norm 1 and $\psi_{i,m} \in W_i$. So, it suffices to show that Eq. 5 holds for unit norm functions in W_i , that the Eq. 5 is tight for such functions, and that given such a function we can construct a Haar-like basis containing it.

The third claim is straightforward to prove. Take $i^* \in \text{branches}$ and let $\psi \in W_{i^*}$ have norm 1. For $i \in \text{branches} \setminus \{i^*\}$, construct an orthonormal basis \mathcal{B}_i for W_i . Similarly, construct for W_{i^*} an orthonormal basis \mathcal{B}_{i^*} containing ψ . This can be done because ψ has norm 1 and is contained in W_{i^*} . Let \mathcal{B} denote the collection $\{\mathbf{1}_\Omega\} \cup \bigcup_{i \in \text{branches}} \mathcal{B}_i$. We claim that \mathcal{B} is a Haar-like basis for V . The only condition of Definition 2 that isn't immediate is the

845 orthogonality of \mathcal{B} . Let $\psi_i \in \mathcal{B}_i \subset \mathcal{B}$. We claim that ψ_i is orthogonal to every other function
 846 in \mathcal{B} . This holds automatically for $\mathbf{1}_{\Omega}$ (by the definition of W_i) and the other members of \mathcal{B}_i
 847 (since \mathcal{B}_i is orthogonal). The remaining case is $\psi_{i'} \in \mathcal{B}_{i'} \subset \mathcal{B}$ with $i \neq i'$. If $i \parallel i'$, then Ω_i
 848 and $\Omega_{i'}$ are disjoint by Lemma 7(b). $\Omega_i \supseteq \text{supp}(\psi_i)$ and $\Omega_{i'} \supseteq \text{supp}(\psi_{i'})$, so ψ_i and $\psi_{i'}$
 849 are then orthogonal. Otherwise, without loss of generality, $i \leq i'$. $i \neq i'$, so in fact $i < i'$.
 850 In particular, there exists some $j \in \text{children}(i)$ such that $j \leq i'$. ψ_i is constant on Ω_j by
 851 Definition 1 (b) and the definition of V_i , and $\psi_{i'} \perp \mathbf{1}_{\Omega_{i'}}$ by the definition of $W_{i'}$. $\Omega_j \supseteq \Omega_{i'}$
 852 by Lemma 7(a), so again $\psi_i \perp \psi_{i'}$.

853 We now return to the claim that Eq. 5 holds and is tight for unit norm functions in W_i .
 854 We begin by bounding the value taken by such functions on a single child of Ω_i . Let $i^* \in$
 855 branches, and let $\Omega_1, \dots, \Omega_k$ be an enumeration of $\text{children}(\Omega_{i^*})$. We seek a solution
 856 to the *

$$\begin{aligned}
 857 \quad & \text{minimize} && -\psi(\Omega_1) && (*) \\
 858 \quad & \text{subject to} && \psi \in V_{i^*} && (*.1) \\
 859 \quad & && \langle \psi, \mathbf{1}_{\Omega_{i^*}} \rangle = 0 && (*.2) \\
 860 \quad & && \langle \psi, \psi \rangle = 1. && (*.3)
 \end{aligned}$$

861 V_{i^*} is in bijection with \mathbb{R}^k , so * can be reformulated as a constrained optimization problem
 862 over Euclidean space. Define $T: V_{i^*} \rightarrow \mathbb{R}^k$ by $T(\phi) = (\phi(\Omega_1), \dots, \phi(\Omega_k))$. Observe that
 863 T is a bijection. Let the objective function $f: \mathbb{R}^k \rightarrow \mathbb{R}$ be given by $f(x) = -x_1$, so that
 864 $f(T(\psi)) = -\psi(\Omega_1)$. Next we must translate each constraint on $\psi \in V_{i^*}$ to a constraint on
 865 $T(\psi) \in \mathbb{R}^k$. Define $h_1, h_2: \mathbb{R}^k \rightarrow \mathbb{R}$ by $h_1(x) = \sum_{i=1}^k A_i x_i$ and $h_2(x) = -1 + \sum_{i=1}^k A_i x_i^2$.

- 866 *.1 Trivially, *.1 holds iff $T(\psi) \in \mathbb{R}^k$.
- 867 *.2 Write A_1, \dots, A_k for the measures $\nu(\Omega_1), \dots, \nu(\Omega_k)$ and A for the sum $A_1 + \dots + A_k$.
- 868 The inner product of ψ and $\mathbf{1}_{\Omega_{i^*}}$ is given by

$$869 \quad \langle \psi, \mathbf{1}_{\Omega_{i^*}} \rangle = \sum_{i=1}^k \nu(\Omega_i) \psi(\Omega_i) = \sum_{i=1}^k A_i T(\psi)_i = h_1(T(\psi)).$$

- 870 *.2 then holds iff $h_1(T(\psi)) = 0$.
- 871 *.3 The inner product of ψ with itself is given by

$$872 \quad \langle \psi, \psi \rangle = \sum_{i=1}^k \nu(\Omega_i) \psi(\Omega_i) \psi(\Omega_i) = \sum_{i=1}^k A_i T(\psi)_i^2 = h_2(T(\psi)) - 1.$$

- 873 *.3 then holds iff $h_2(T(\psi)) = 0$.

874 * can therefore be rewritten

$$\begin{aligned}
 875 \quad & \text{minimize} && f(x) && (\dagger) \\
 876 \quad & \text{subject to} && x \in \mathbb{R}^k && (\dagger.1) \\
 877 \quad & && h_1(x) = 0 && (\dagger.2) \\
 878 \quad & && h_2(x) = 0. && (\dagger.3)
 \end{aligned}$$

879 The Lagrangian function $L_f: \mathbb{R}^k \times \mathbb{R}^2 \rightarrow \mathbb{R}$ is given by

$$881 \quad L_f(x, \lambda) = f(x) + \lambda_1 h_1(x) + \lambda_2 h_2(x).$$

882 Its gradient with respect to x is given by

883
$$\nabla_x L_f(x, \lambda) = (-1, 0, \dots, 0) + \lambda_1(A_1, \dots, A_k) + 2\lambda_2(A_1x_1, \dots, A_kx_k).$$

884 We will use the method of Lagrange multipliers to find the global minimum of †. First, we
 885 will apply a necessary condition to find two candidate local minima. Then, we will apply a
 886 sufficient condition to show that one of the two is the global minimum.

887 The gradients of the feasibility constraints are given by

888
$$\nabla h_1(x) = (A_1, \dots, A_k) \quad \text{and} \quad \nabla h_2(x) = 2(A_1x_1, \dots, A_kx_k).$$

889 Let \mathfrak{S} denote the feasible set of †. We claim that ∇h_1 and ∇h_2 are linearly independent on \mathfrak{S} .
 890 Let $x \in \mathfrak{S}$. Each of A_1, \dots, A_k is positive. So, in order for $h_1(x)$ to be zero, x must either be $\mathbf{0}$
 891 or have at least one positive and at least one negative component. $h_2(\mathbf{0}) = -1$, so x cannot be
 892 $\mathbf{0}$. x therefore has at least one positive and at least one negative component. $\nabla h_2(x)$ therefore
 893 likewise has at least one positive and at least one negative component. All the components
 894 of $\nabla h_1(x)$, though, are positive. In particular, $\nabla h_1(x)$ and $\nabla h_2(x)$ are linearly independent.

895 Suppose that x^* is a local minimum of †. Because f, h_1 , and h_2 are continuously differ-
 896 entiable and ∇h_1 and ∇h_2 are linearly independent on \mathfrak{S} , there exist Lagrange multipliers
 897 $\lambda^* \in \mathbb{R}^2$ such that the gradient with respect to x of the Lagrangian at (x^*, λ^*) is zero [48,
 898 Proposition 3.1.1]. That is,

899
$$0 = -1 + A_1\lambda_1^* + 2A_1\lambda_2^*x_1^* \tag{6}$$

900
$$0 = 0 + A_i\lambda_1^* + 2A_i\lambda_2^*x_i^* \quad 2 \leq i \leq k. \tag{7}$$

901 If $\lambda_2^* = 0$, then $\lambda_1^* = 0$ by Eq. 7, contradicting Eq. 6. λ_2^* is therefore nonzero and so Eq. 7
 902 can be simplified to $x_i^* = -\lambda_1^*/2\lambda_2^*$ for all $2 \leq i \leq k$. Substituting into †.2, we obtain an
 903 expression for x_1^* :

904
$$0 = h_1(x^*) = A_1x_1^* + \sum_{i=2}^k A_ix_i^* = A_1x_1^* - (A - A_1)\frac{\lambda_1^*}{2\lambda_2^*}$$

 905
$$x_1^* = \left(\frac{A}{A_1} - 1\right)\frac{\lambda_1^*}{2\lambda_2^*}. \tag{8}$$

906 We next solve for λ_1^* using Eq. 6:

907
$$0 = -1 + A_1\lambda_1^* + 2A_1\lambda_2^*\left(\frac{A}{A_1} - 1\right)\frac{\lambda_1^*}{2\lambda_2^*} = -1 + A_1\lambda_1^* + (A - A_1)\lambda_1^*$$

 908
$$\lambda_1^* = 1/A. \tag{9}$$

909 Next, apply †.3.

910
$$1 = A_1(x_1^*)^2 + \sum_{i=2}^k A_i(x_i^*)^2 = A_1\left[\left(\frac{A}{A_1} - 1\right)\frac{\lambda_1^*}{2\lambda_2^*}\right]^2 + (A - A_1)\left[-\frac{\lambda_1^*}{2\lambda_2^*}\right]^2$$

 911
$$[2\lambda_2^*]^2 = \left[A_1\left(\frac{A}{A_1} - 1\right)\right]^2 + (A - A_1)[\lambda_1^*]^2 = \left[\frac{A^2}{A_1} - A\right][\lambda_1^*]^2 = \left[\frac{1}{A_1} - \frac{1}{A}\right]$$

 912
$$\lambda_2^* = \pm \frac{1}{2}\sqrt{\frac{1}{A_1} - \frac{1}{A}} \tag{10}$$

913 We will write $\lambda_{+,2}^*$ for the positive square root and $\lambda_{-,2}^*$ for the negative square root. x_+^* and
 914 x_-^* will denote the corresponding candidate local minima. Equations 7 and 10 together yield
 915 an expression for $x_{\pm,i}^*$: for $2 \leq i \leq k$,

916
$$0 = 0 + \frac{A_i}{A} \pm \frac{2}{2}A_i\sqrt{\frac{1}{A_1} - \frac{1}{A}}x_{\pm,i}^*$$

 917
$$x_{\pm,i}^* = \mp \frac{1/A}{\sqrt{1/A_1 - 1/A}} = \mp \sqrt{\frac{A_1/A}{A - A_1}}.$$

918 Similarly, Eqs. 8, 9, and 10 yield an expression for $x_{\pm,1}^*$:

919
$$x_{\pm,1}^* = \left(\frac{A}{A_1} - 1\right) \frac{1/A}{\pm \frac{1}{2} \sqrt{1/A_1 - 1/A}} = \pm \left(\frac{A}{A_1} - 1\right) \sqrt{\frac{A_1/A}{A - A_1}}.$$

920 The candidate local minima of † are then

921
$$x_{\pm}^* = \pm \sqrt{\frac{A_1/A}{A - A_1}} \left(\frac{A}{A_1} - 1, -1, \dots, -1\right).$$

922 We claim that x_+^* is a local minimum with Lagrange multipliers $(\lambda_1^*, \lambda_{+,2}^*)$. Observe that f ,
 923 h_1 , and h_2 are twice continuously differentiable. As shown above, $\nabla_x L_f(x_+^*, (\lambda_1^*, \lambda_{+,2}^*)) =$
 924 0 . Because $x_+^* \in \mathfrak{S}$,

925
$$\nabla_{\lambda} L_f(x_+^*, (\lambda_1^*, \lambda_{+,2}^*)) = (h_1(x_+^*), h_2(x_+^*)) = \mathbf{0}.$$

926 The Hessian with respect to x of the Lagrangian is given by

927
$$\nabla_{xx}^2 L_f(x, \lambda) = 2\lambda_2 \begin{bmatrix} A_1 & & \\ & \ddots & \\ & & A_k \end{bmatrix}. \tag{11}$$

928 Each of A_1, \dots, A_k is positive, so $\nabla_{xx}^2 L_f$ is symmetric positive definite at (x, λ) if $\lambda_2 > 0$.
 929 $\lambda_{+,2}^* > 0$, so x_+^* is a local minimum of † [48, Proposition 3.2.1].

930 Denote by (‡) the problem of minimizing $-f$ with the constraints of †. We claim that x_-^*
 931 is a local minimum of (‡) with Lagrange multipliers $(-\lambda_1^*, -\lambda_{-,2}^*)$. $-f, h_1$, and h_2 are twice
 932 continuously differentiable. Let $L_{-f}: \mathbb{R}^k \times \mathbb{R}^2 \rightarrow \mathbb{R}$ be the Lagrangian function:

933
$$L_{-f}(x, \lambda) = -f(x) + \lambda_1 h_1(x) + \lambda_2 h_2(x).$$

934 L_{-f} is related to the Lagrangian L_f of † as follows:

935
$$\begin{aligned} \nabla_x L_{-f}(x, \lambda) &= -\nabla_x f(x) + \lambda_1 \nabla_x h_1(x) + \lambda_2 \nabla_x h_2(x) \\ 936 &= -\nabla_x f(x) - (-\lambda_1) \nabla_x h_1(x) - (-\lambda_2) \nabla_x h_2(x) = -\nabla_x L_f(x, -\lambda). \end{aligned}$$

937 As a result,

938
$$\nabla_x L_{-f}(x_-^*, (-\lambda_1^*, -\lambda_{-,2}^*)) = -\nabla_x L_f(x_-^*, (\lambda_1^*, \lambda_{-,2}^*)) = 0.$$

939 Since $x_-^* \in \mathfrak{S}$,

940
$$\nabla_{\lambda} L_{-f}(x_-^*, (-\lambda_1^*, -\lambda_{-,2}^*)) = (h_1(x_-^*), h_2(x_-^*)) = \mathbf{0}.$$

941 Because $\nabla_x L_{-f}(x, \lambda) = -\nabla_x L_f(x, -\lambda)$, $\nabla_{xx}^2 L_{-f}(x, \lambda) = -\nabla_{xx}^2 L_f(x, -\lambda)$. Referring to
 942 Eq. 11, we see that $\nabla_{xx}^2 L_f$ is symmetric negative definite at $(x_-^*, (\lambda_1^*, \lambda_{-,2}^*))$, since $\lambda_{-,2}^* < 0$.
 943 $\nabla_{xx}^2 L_{-f}$ is then symmetric positive definite at $(x_-^*, (-\lambda_1^*, -\lambda_{-,2}^*))$. We conclude that x_-^* is
 944 a local minimum of (‡) [48, Proposition 3.2.1].

945 As a local minimum of (‡), x_-^* is a local maximum of †. In particular, x_+^* is the only
 946 local minimum of the latter †. \mathfrak{S} is compact, so x_+^* must be the global minimum. The global
 947 minimum of * is therefore the function $\psi^* \in V_i^*$ defined by

948
$$\psi^* = T^{-1}(x_+^*) = \sqrt{\frac{A_1/A}{A - A_1}} \left[\left(\frac{A}{A_1} - 1\right) \mathbf{1}_{\Omega_1} - \sum_{i=2}^k \mathbf{1}_{\Omega_i} \right].$$

949 If $A_1 \leq A/2$, then

950
$$\|\psi^*\|_{C^0} = \sqrt{\frac{A_1/A}{A-A_1}} \left(\frac{A}{A_1} - 1 \right) = \sqrt{\frac{A-A_1}{AA_1}} = \sqrt{\frac{1}{A_1} - \frac{1}{A}} \leq \sqrt{\frac{1}{\min_{1 \leq i \leq k} A_i} - \frac{1}{A}}.$$

951 Eq. 5 is therefore respected in this case. The Eq. 5 is tight if A_1 is minimal. If instead $A_1 \geq$
 952 $A/2$, then

953
$$\|\psi^*\|_{C^0} = \sqrt{\frac{A_1/A}{A-A_1}} = \sqrt{\frac{A-(A-A_1)}{A(A-A_1)}} = \sqrt{\frac{1}{A-A_1} - \frac{1}{A}}.$$

954 $A - A_1 \geq \min_{1 \leq i \leq k} A_i$, so the Eq. 5 again holds. The Eq. 5 is tight if Ω_i^* has two children,
 955 of which Ω_1 is the larger, so that $A - A_1 = A_2 = \min_{1 \leq i \leq k} A_i$. □

956 **References**

957 1. Robinson, A.H., Cherry, C.: Results of a prototype television bandwidth compression scheme. Proc. IEEE
 958 **55**(3), 356–364 (1967)

959 2. Bradley, Stevan D.: Optimizing a scheme for run length encoding. Proc. IEEE **57**(1), 108–109 (1969)

960 3. Hauck, Edward L.: Data compression using run length encoding and statistical encoding, December 2
 961 (1986). US Patent 4,626,829

962 4. Ziv, J., Lempel, A.: A universal algorithm for sequential data compression. IEEE Trans. Inf. Theory **23**(3),
 963 337–343 (1977)

964 5. Pavlov, Igor: LZMA specification (draft), (June 2015)

965 6. Nevill-Manning, C.G., Witten, I.H.: Identifying hierarchical structure in sequences: a linear-time algo-
 966 rithm. J. Artif. Intell. Res. **7**, 67–82 (1997)

967 7. Mandagere, N., Zhou, P., Smith, MA., Uttamchandani, S.: Demystifying data deduplication. In: Proceed-
 968 ings of the ACM/IFIP/USENIX Middleware '08 Conference Companion, pp. 12–17, (2008)

969 8. Manber, U.: Finding similar files in a large file system. In: USENIX Winter 1994 Technical Conference
 970 Proceedings, vol. 94, pp. 1–10, (1994)

971 9. Xia, Wen, J., Hong, F., Dan, H., Yu: S.: A similarity-locality based near-exact deduplication scheme
 972 with low ram overhead and high throughput. In: Proceedings of the 2011 USENIX Annual Technical
 973 Conference, pp. 26–30, (2011)

974 10. Wallace, G.K.: The JPEG still picture compression standard. IEEE Trans. Consum. Electron. **38**(1), 18–34
 975 (1992)

976 11. Grgic, S., Kers, K., Grgic, M.: Image compression using wavelets. In: Proceedings of the IEEE Interna-
 977 tional Symposium on Industrial Electronics. ISIE '99, vol. 1, pp. 99–104, (1999)

978 12. Marcellin, M.W., Gormish, M.J., Bilgin, A., Boliek, M.P.: An overview of JPEG-2000. In: Proceedings
 979 DCC 2000. Data Compression Conference, pp. 523–541, (2000)

980 13. Tang, Xiaoli, Pearlman, William A: Lossy-to-lossless block-based compression of hyperspectral volu-
 981 metric data. In: 2004 International Conference on Image Processing. ICIP '04., vol. 5, pp. 3283–3286.
 982 IEEE, (2004)

983 14. Lindstrom, Peter: Fixed-rate compressed floating-point arrays. IEEE Trans. Visual Comput. Gr. **20**(12),
 984 2674–2683 (2014)

985 15. Li, Shaomeng, Jaroszynski, Stanislaw, Pearse, Scott, Orf, Leigh, Clyne, John: VAPOR: a visualization
 986 package tailored to analyze simulation data in earth system science. Atmosphere **10**(9), 488 (2019)

987 16. Ainsworth, Mark, Tugluk, Ozan, Whitney, Ben, Klasky, Scott: Multilevel techniques for compression and
 988 reduction of scientific data—quantitative control of accuracy in derived quantities. SIAM J. Sci. Comput. **41**(4),
 989 A2146–A2171 (2019)

990 17. Austin, W., Ballard, G., Kolda, T.G.: Parallel tensor compression for large-scale scientific data. In: 2016
 991 IEEE International Parallel and Distributed Processing Symposium (IPDPS), pp. 912–922, (2016)

992 18. Ballester-Ripoll, Rafael, Lindström, Peter, Pajarola, Renato: TTHRESH: tensor compression for multi-
 993 dimensional visual data. IEEE Trans. Visual Comput. Gr. **26**(9), 2891–2903 (2020)

994 19. Wu, Qing, Xia, Tian, Yu, Yizhou: Hierarchical tensor approximation of multidimensional images. In:
 995 2007 IEEE International Conference on Image Processing, vol. 4, pp. 49–52. IEEE, (2007)

996 20. Jiang, W.W., Kiang, S.Z., Hakim, N.Z., Meadows, H.E.: Lossless compression for medical imaging
 997 systems using linear/nonlinear prediction and arithmetic coding. In: ISCAS '93, IEEE International Sym-
 998 posium on Circuits and Systems, vol. 1, pp. 283–286, (1993)

- 999 21. Lindstrom, Peter, Isenburg, Martin: Fast and efficient compression of floating-point data. *IEEE Trans. Visual Comput. Gr.* **12**(5), 1245–1250 (2006)
- 1000 22. Roelofs, Greg: PNG: The Definitive Guide. O'Reilly Media, Sebastopol (1999)
- 1001 23. Bautista Gomez, L.A., Cappello, F: Improving floating point compression through binary masks. In: 2013
- 1002 IEEE International Conference on Big Data, pp. 326–331, (2013)
- 1003 24. Di, S., Cappello, F: Fast error-bounded lossy HPC data compression with SZ. In: 2016 IEEE 30th
- 1004 International Parallel and Distributed Processing Symposium, Chicago, IL, USA, pp. 730–739 (2016).
- 1005 IEEE
- 1006 25. Tao, D., Di, S., Chen, Z., Cappello, F: Significantly improving lossy compression for scientific data
- 1007 sets based on multidimensional prediction and error-controlled quantization. In: 2017 IEEE International
- 1008 Parallel and Distributed Processing Symposium (IPDPS), pp. 1129–1139, Orlando, FL, USA, (2017).
- 1009 IEEE
- 1010 26. Ainsworth, Mark, Tugluk, Ozan, Whitney, Ben, Klasky, Scott: Multilevel techniques for compression and
- 1011 reduction of scientific data—the unstructured case. *SIAM J. Sci. Comput.* **42**(2), A1402–A1427 (2020)
- 1012 27. Shuman, D.I., Narang, S.K., Frossard, P., Ortega, A., Vandergheynst, P.: The emerging field of signal
- 1013 processing on graphs: extending high-dimensional data analysis to networks and other irregular domains.
- 1014 *IEEE Signal Process. Mag.* **30**(3), 83–98 (2013)
- 1015 28. Avena, Luca, Castell, Fabienne, Gaudillière, Alexandre, Mélot, Clothilde: Intertwining wavelets or mul-
- 1016 ti-resolution analysis on graphs through random forests. *Appl. Comput. Harmon. Anal.* **48**(3), 949–992
- 1017 (2020)
- 1018 29. Coifman, Ronald R., Maggioni, M.: Diffusion wavelets. *Appl. Comput. Harmonic Anal.* **21**(1), 53–94
- 1019 (2006)
- 1020 30. Hammond, David K., Vandergheynst, Pierre, Gribonval, Rémi.: Wavelets on graphs via spectral graph
- 1021 theory. *Appl. Comput. Harmon. Anal.* **30**(2), 129–150 (2011)
- 1022 31. Murtagh, Fionn: The Haar wavelet transform of a dendrogram. *J. Classif.* **24**(1), 3–32 (2007)
- 1023 32. Lee, Ann B., Nadler, Boaz, Wasserman, Larry: Treelets—an adaptive multi-scale basis for sparse unordered
- 1024 data. *Ann. Appl. Stat.* **2**(2), 435–471 (2008)
- 1025 33. Elisha, Oren, Dekel, Shai: Wavelet decompositions of random forests: smoothness analysis, sparse approx-
- 1026 imation and applications. *J. Mach. Learn. Res.* **17**(1), 6952–6989 (2016)
- 1027 34. Salloum, Maher, Fabian, Nathan D., Hensinger, David M., Lee, Jina, Allendorf, Elizabeth M., Bhagatwala,
- 1028 Ankit, Blaylock, Myra L., Chen, Jacqueline H., Templeton, Jeremy A., Tezaur, Irina: Optimal compressed
- 1029 sensing and reconstruction of unstructured mesh datasets. *Data Sci. Eng.* **3**(1), 1–23 (2018)
- 1030 35. Bender, EA., Williamson, SG: Lists, decisions and graphs. S. Gill Williamson, (2010)
- 1031 36. Gavish, Matan, Nadler, Boaz, Coifman, Ronald R: Multiscale wavelets on trees, graphs and high dimensional
- 1032 data: theory and applications to semi supervised learning. In: ICML, pp. 367–374, (2010)
- 1033 37. Shapiro, Jerome M.: Embedded image coding using Zerotrees of wavelet coefficients. *IEEE Trans. Signal*
- 1034 *Process.* **41**(12), 3445–3462 (1993)
- 1035 38. Jarlskog, Cecilia: A recursive parametrization of unitary matrices. *J. Math. Phys.* **46**(10), 103508 (2005)
- 1036 39. Shilov, Georgi E., Silverman, Richard A., et al.: Elementary real and complex analysis. Courier Corpora-
- 1037 tion, Chelmsford (1996)
- 1038 40. Bentley, Jon Louis: Multidimensional binary search trees used for associative searching. *Commun. ACM*
- 1039 **18**(9), 509–517 (1975)
- 1040 41. LeCun, Yann, Bottou, Léon., Bengio, Yoshua, Haffner, Patrick: Gradient-based learning applied to document
- 1041 recognition. *Proc. IEEE* **86**(11), 2278–2324 (1998)
- 1042 42. Lepelaars, Carlo: 97% on MNIST with a single decision tree (+ t-SNE). <https://www.kaggle.com/code/carlolepelaars/97-on-mnist-with-a-single-decision-tree-t-sne>, (November 2019). Version 26
- 1043 43. Halko, Nathan, Martinsson, Per-Gunnar., Tropp, Joel A.: Finding structure with randomness: probabilistic
- 1044 algorithms for constructing approximate matrix decompositions. *SIAM Rev.* **53**(2), 217–288 (2011)
- 1045 44. Pedregosa, Fabian, Varoquaux, Gaël., Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel,
- 1046 Olivier, Blondel, Mathieu, Prettenhofer, Peter, Weiss, Ron, Dubourg, Vincent, Vanderplas, Jake, Passos,
- 1047 Alexandre, Cournapeau, David, Brucher, Matthieu, Perot, Matthieu, Duchesnay, Edouard: Scikit-learn:
- 1048 machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
- 1049 45. Linderman, George C., Rachh, Manas, Hoskins, Jeremy G., Steinerberger, Stefan, Kluger, Yuval: Fast
- 1050 interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat. Methods* **16**(3),
- 1051 243–245 (2019)
- 1052 46. Poličar, Pavlin G., Stražar, Martin, Zupan, Blaž: openTSNE: a modular Python library for t-SNE dimensional-
- 1053 ity reduction and embedding. *bioRxiv*, (2019)
- 1054 47. Sattath, Shmuel, Tversky, Amos: Additive similarity trees. *Psychometrika* **42**(3), 319–345 (1977)
- 1055 48. Bertsekas, Dimitri: *Nonlinear Programming*, 2nd edn. Athena Scientific, Belmont (1999)
- 1056
- 1057

1058 **Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and
1059 institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

uncorrected proof