CART and Best-Ortho-Basis: A Connection

David L. Donoho
Statistics
Stanford University

July 1995

Abstract

We study what we call "Dyadic CART" – a method of nonparametric regression which constructs a recursive partition by optimizing a complexity-penalized sum of squares, where the optimization is over all recursive partitions arising from midpoint splits. We show that the method is adaptive to unknown degree of anisotropic smoothness. Specifically, consider the "mixed smoothness" classes consisting of bivariate functions $f(x_1, x_2)$ whose finite difference of distance h in direction i is bounded in L^p norm by Ch^{δ_i}, $i = 1, 2$. We show that Dyadic CART, with an appropriate complexity penalty parameter $\lambda \sim \sigma^2 \cdot Const \cdot \log(n)$, is within logarithmic terms of minimax over every mixed smoothness class $0 < C < \infty$, $0 < \delta_1, \delta_2 \leq 1$.

The proof shows that Dyadic CART is identical to a certain adaptive Best-Ortho-Basis Algorithm based on the library of all anisotropic Haar bases. Then it applies empirical basis selection ideas of Donoho and Johnstone (1994). The basis empirically selected by Dyadic CART is shown to be nearly as good as a basis ideally adapted to the underlying f. The risk of estimation in an ideally adapted anisotropic Haar basis is shown to be comparable to the minimax risk over mixed smoothness classes.

Underlying the success of this argument is harmonic analysis of mixed smoothness classes. We show that for each mixed smoothness class there is an anisotropic Haar basis which is a best orthogonal basis for representing that smoothness class; the basis is optimal not just within the library of anisotropic Haar bases, but among all orthogonal bases of $L^2[0, 1]^2$.

Key Words and Phrases. Wavelets, Mixed Smoothness, Anisotropic Haar Basis, Best Orthogonal Basis, Minimax Estimation, Spatial Adaptation, Oracle Inequalities.

Acknowledgements. This paper was stimulated by some interesting conversations about CART and wavelets which the author had with Joachim Engel at Oberwolfach, March 1995. It is a pleasure also to acknowledge conversations with R.R. Coifman and I.M. Johnstone.

The author is also at U.C. Berkeley (on leave). This research was partially supported by NSF DMS-92-09130.

Thanks to Helen Tombropoulos for an efficient and enthusiastic typing job.

1 Introduction

The CART methodology of tree-structured adaptive non-parametric regression [4] has been widely used in statistical data analysis since its inception more than a decade ago. Built around ideas of recursive partitioning, it develops, based on an analysis of noisy data, a piecewide constant reconstruction, where the pieces are terminal nodes of a data-driven recursive partition.

The Best-Ortho-Basis methodology of adaptive time-frequency analysis [5] has, more recently, caught the interest of a wide community of applied mathematicians and signal processing engineers. Based on ideas of recursive-partitioning of the time-frequency plane, it develops, from an analysis of a given signal, a segmented basis, where the segments are terminal nodes in a data-driven recursive segmentation of the time axis.

Both methods are concerned with recursive dyadic segmentation; therefore trees and tree pruning are key data structures and underlying algorithms in both areas. In addition, there is a mathematical connection between the areas.

Sudeshna Adak, currently a graduate student at Stanford University, has pointed out that central algorithms in the two subjects are really the same: namely the optimal pruning algorithm in Theorem 10.7, page 285, in the CART book [4] and in the Proposition on page 717, in the Best-Basis paper [6]. Both theorems assert that, given a function $\mathcal{E}(T)$ which assigns numerical values to a binary tree, and its subtrees, and supposing that the function obeys a certain additivity property, the optimal subtree is obtained by breadth-first, bottom-up pruning of the complete tree.

On the other hand, the subjects are also different, since in one case (CART) one is searching for an optimal *function* on a multidimensional cartesian product domain, and in the other one (BOB) is searching for an optimal orthogonal *basis* for the vector space of $1-d$ signals of length n.

This paper will exhibit a precise connection between CART and BOB in a specific setting — where one is seeking an optimal function/basis built from rectangular blocks on a product domain. In this setting we show that certain specific variants of the two apparently different methodologies lead to identical fast algorithms and identical solutions.

1.1 An implication

The connection between CART and Best-Basis affords new insights about recursive partitioning methods. Recently, Donoho and Johnstone have investigated the use of adaptively chosen bases for noise removal [11]. They have developed so-called oracle inequalities which show that certain schemes for basis selection in the presence of noisy data will work well. By adapting such ideas from the Best-Basis setting to the CART setting, we are able to establish new results on the performance of optimal dyadic recursive partitioning. In particular, we are able to show that such methods can be nearly-minimax simultaneously over a wide range of mixed smoothness spaces.

We assume observations of the form

$$y(i_1,i_2) = \bar{f}(i_1,i_2) + \sigma z(i_1,i_2) \quad 0 \le i_1, i_2 < n, \tag{1.1}$$

where n is dyadic (a integral power of 2), z_{i_1,i_2} is a white Gaussian noise, and $\sigma > 0$ is a noise level. We assume the observations are related to the underlying f by cell averaging;

$$\bar{f}(i_1,i_2) = Ave\{f | [i_1/n,(i_1+1)/n) \times [i_2/n,(i_2+1)/n)\}. \tag{1.2}$$

Our goal is to recover the de-noised cell averages with small mean squared error $E\|\hat{f} - \bar{f}\|_{\ell^2}^2 = E\sum_{i_1,i_2}(\hat{f}(i_1,i_2) - \bar{f}(i_1,i_2))^2$. About f we will assume that it belongs to a certain class $\mathcal{F}$, and we will compare performance of estimates with the best mean-squared error available uniformly over the class $\mathcal{F}$ – i.e. the minimax risk:

$$M^*(\sigma, n; \mathcal{F}) = \inf_{\hat{f}(\cdot)} \sup_{f \in \mathcal{F}} E\|\hat{f}(y) - \bar{f}\|_{\ell^2}^2. \tag{1.3}$$

For our $\mathcal{F}$ we consider mixed smoothness classes $\mathcal{F}_p^{\delta_1,\delta_2}(C)$ consisting of functions on $[0,1]^2$ obeying $\|D_h^1 f\|_p \le Ch^{\delta_1}$, $\|D_h^2 f\|_p \le Ch^{\delta_2}$, for all $h \in (0,1)$, where D_h^i denotes the finite difference of distance h in direction i. We let $\mathcal{MS}$ denote the scale of all such classes, where $0 < p < \infty$, $0 < \delta_1, \delta_2 \le 1$ and $0 < C < \infty$.

Our main result:

Theorem 1.1 *Dyadic CART (defined in section 2 below), with the specific complexity penalty $\lambda = \lambda(\sigma, \log_e(n))$ defined in section 7 below ($\lambda \asymp \sigma^2 \log_e(n)$), comes within logarithmic factors of minimax over each functional class $\mathcal{F}_p^{\delta_1,\delta_2}(C)$, $0 < \delta_1, \delta_2 \le 1$, $C > 0$, $p \in (0,\infty)$. If $\hat{f}^{*,\lambda}$ denotes the dyadic CART estimator*

$$\sup_{\mathcal{F}} E\|\hat{f}^{*,\lambda} - \bar{f}\|_{\ell^2}^2 \le Const(\delta_1, \delta_2, p) \cdot \log(n) \cdot M^*(\sigma, n; \mathcal{F}) \qquad \text{as } n \to \infty. \tag{1.4}$$

for each $\mathcal{F} \in \mathcal{MS}$.

In short, the estimator behaves nearly as well over any class in the scale $\mathcal{MS}$ as one could achieve *knowing precisely* which smoothness class were true. However, the construction of the optimal recursive partitioning estimator requires no knowledge of which smoothness class might actually be the case. (Indeed, we are unaware of any previous literature suggesting a connection between such smoothness classes and CART).

We remark that no similar adaptivity is possible using standard isotropic wavelets or isotropic Fourier analysis. This illustrates a theoretical benefit to using recursive partitioning in a setting where objects may possess different degrees of smoothness in different directions.

1.2 Plan of the Paper

In sections 2 through 6 we develop the connection between CART methods and Best Basis methods. Section 2 defines Dyadic CART and describes its fast algorithm. Section 3 defines a library of Anisotropic Haar Bases and describes a fast algorithm for finding a Best Anisotropic Haar basis from given data, where "best" is defined in the Coifman-Wickerhauser sense. In Sections 4 and 5, building on an insight of Joachim Engel (1994), we point out that, with traditional choices of entropy, Best Ortho Basis is different from CART, but that with a special *Hereditary Entropy* the two methods are the same.

In sections 7 and 8 we discuss ideas first developed in the best basis setting. Section 7 develops oracle inequalities, which show how to select a basis empirically from noisy data to yield a basis that is nearly as good as the ideal basis which could be designed based on noiseless data. Section 8 describes the best-basis problem for mixed smoothness classes, and shows that a certain kind of anisotropic Haar basis is, in a certain sense, a best basis.

In section 9, building on sections 7 and 8, we showthat a certain best-basis de-noising technique introduced by Donoho and Johnstone [11] — which is different from CART — is nearly minimax over the scale of mixed smoothness classes. Section 10 establishes our main result for CART by comparing the CART estimator with this best-basis de-noising method, and showing that the two estimates have comparable performance over mixed smoothness spaces. Section 11 discusses comparisons and generalizations.

2 Dyadic CART

We change notation slightly from (1.1). We suppose we observe noisy 2-dimensional data on a regular square $n \times n$ array of "pixels",

$$y(i_1, i_2) = f(i_1, i_2) + \sigma \cdot z(i_1, i_2), \quad 0 \leq i_1, i_2 < n, \tag{2.1}$$

where (in a change from the last section) f is the object of interest – an n-by-n array – and z is a standard Gaussian white noise (i.i.d. $N(0,1)$). We also introduce a fruitful abuse of notation. We write $[0, n)$ for the *discrete interval* $\{0, \cdots, n-1\}$. Thus $[0, n)^2$ is a discrete square, etc. Here and below we also write $\mathbf{i} = (i_1, i_2)$, so $y(\mathbf{i}) = f(\mathbf{i}) + \sigma z(\mathbf{i})$, for $\mathbf{i} \in [0, n)^2$ is an equivalent form of (2.1). Finally, we use the variable $N = n^2$ to stand for the cardinality of the n-by-n array y.

In this setting, the CART methodology constructs a piecewise constant estimator $\hat{f}$ of f; data-adaptively, it builds a partition $\mathcal{P}$ of $[0, n)^2$ and finds $\hat{f}$ by the rule

$$\hat{f}(\mathbf{i}|\mathcal{P}) = Ave\{y \mid R(\mathbf{i}; \mathcal{P})\} \tag{2.2}$$

where $R(\mathbf{i}; \mathcal{P})$ denotes the rectangle of the partition $\mathcal{P}$ containing $\mathbf{i}$.

2.1 Optimal Dyadic CART

There are several variants of CART, depending on the procedure used to construct the partitition $\mathcal{P}$. In this paper, we are only interested in *optimal* (non-greedy) dyadic recursive partitioning. With an acknowledged risk of misunderstanding, we call this *dyadic CART*. We define terms.

Dyadic Partitioning. Starting from the trivial partition $\mathcal{P}_0 = \{[0, n)^2\}$ we may generate new partitions by splitting $[0, n)^2$ into two pieces either vertically or horizontally, yielding either the partitition $\{[0, n/2) \times [0, n), [n/2, n) \times [0, n)\}$ or $\{[0, n) \times [0, n/2), [0, n) \times [n/2, n)\}$. We can apply this splitting recursively, generating other partitions. Thus, let $P = \{R_1, \cdots, R_k\}$ be a partition and let R stand for one of the rectangles in the partition. We can create a new partition by splitting R in half horizontally or vertically. If $R = [a, b) \times [c, d)$ then let $R^{1,0}$ and $R^{1,1}$ denote the results of horizontal splitting, i.e.

$$\begin{aligned} R^{1,0} &= [a_1(a+b)/2) \times [a, d], \\ R^{1,1} &= [(a+b)/2, b) \times [c, d]; \end{aligned}$$

while we let $R^{2,0}$ and $R^{2,1}$ denote the results of vertical splitting

$$\begin{aligned} R^{2,0} &= [a, b) \times [c, (c+d)/2), \\ R^{2,1} &= [a, b) \times [(c+d)/2, d). \end{aligned}$$

Note that if $b = a + 1$ or $d = c + 1$ then horizontal/vertical splitting is not possible: only nonempty rectangles are allowed.

As an example, if we split vertically the rectangle $R = R_\ell$, say, we produce the $k+1$-element partition $\{R_1, \cdots, R_{\ell-1}, R_\ell^{2,0}, R_\ell^{2,1}, R_{\ell+1}, \cdots, R_k\}$.

A *recursive dyadic partition* is any partition reachable by successive application of these rules.

Optimal Partitions. CART is often used to refer to "greedy growing" followed by "optimal pruning", where the partition $\mathcal{P}$ is constructed in a heuristic, myopic fashion. For the purposes of this paper, we consider instead the use of optimizing partitions, where the dyadic partition $\mathcal{P}$ is constructed as the optimum of the *complexity-penalized residual sum of squares.* Thus, with

$$CPRSS(\mathcal{P}, \lambda) = \|y - \hat{f}(\cdot|\mathcal{P})\|_{\ell^2_N}^2 + \lambda \#(\mathcal{P}), \tag{2.3}$$

what we will call (again in perhaps a slight abuse of nomenclature) dyadic CART seeks the partition

$$\hat{\mathcal{P}}_\lambda = \underset{\mathcal{P}}{\operatorname{argmin}}\, CPRSS(\mathcal{P}, \lambda) \tag{2.4}$$

The idea of using globally optimal partitions is covered in passing in [4, Chapter 10]. For the moment we let λ be a free parameter; in section 7 below we will propose a specific choice.

Dyadic CART differs from what is usually called CART, in that dyadic CART can split rectangles only in half, while general CART can split rectangles in all proportions. While the extra flexibility of general CART may be useful, this flexibility is sufficient to make the finding of an exactly optimal partition unwieldy. Dyadic CART allows a more limited range of possible partitions, which makes it possible to find an optimal partition in order $O(N)$ time.

2.2 Fast Optimal Partitioning

To describe the algorithm, we introduce some notation.

Rectangles. We use generically I to denote dyadic intervals, i.e. intervals $I = [a, b)$ with $a = n \cdot k/2^j$ and $b = n \cdot (k+1)/2^j$ with $n \geq 2^j$ and $0 \leq k < 2^j$. We use R to denote dyadic rectangles, i.e., rectangles $I_1 \times I_2$.

Parents and Siblings. Two dyadic rectangles are *siblings* if their union is a dyadic rectangle. This is equivalent to saying that we can write either

$$R_i = I_i \times I_0 \quad i = 1, 2, \tag{2.5}$$

or

$$R_i = I_0 \times I_i \quad i = 1, 2, \tag{2.6}$$

where I_0, I_1, I_2 are dyadic intervals and

$$I_1 = [n \cdot 2k/2^j, n \cdot (2k+1)/2^j), \quad I_2 = [n \cdot 2(k+1)/2^j, n \cdot (2k+2)/2^j), \tag{2.7}$$

with $0 \leq k < 2^{j-1}, 0 \leq j < \log_2(n) - 1$. A pair satisfying (2.5) is a pair of *horiz-sibs*; a pair satisfying (2.6) is a pair of *vert-sibs*.

The union of two siblings is the parent rectangle. Each rectangle generally has two siblings — a vert-sib and a horiz-sib — and two parents — a vert-parent and an horiz-parent. Parents generally have two sets of children: a pair of horiz-kids and a pair of vert-kids. In extreme cases a rectangle may have only a vert-sib (if it is very wide, such as $[0,n) \times [0,n/2))$, or only an horiz-sib (if it is very tall, such as $[0,n/2) \times [0,n))$. In some cases a rectangle may have only vert-kids (if it is very narrow, such as $[0,1) \times [0,n/2))$ or only horiz-kids (if it is very short, such as $[0,n/2) \times [0,1))$.

Inheritance. CPRSS has an "inheritance property" which we more easily see by taking a general point of view. Let $CART(R)$ denote the problem of finding the optimal partition for *just the data falling in the dyadic rectangle* R:

$$(CART(R)) \qquad \hat{\mathcal{P}}(R) = \text{argmin}\, \|y - \hat{f}(\cdot|\mathcal{P}(R))\|^2_{\ell^2(R)} + \lambda \cdot \#(\mathcal{P}(R)).$$

Here $\mathcal{P}(R)$ denotes a recursive dyadic partition of R, and $\|\cdot\|^2_{\ell^2(R)}$ refers to the sum-of-squares only of data falling in the rectangle R.

Here is the inheritance property of optimal partitions. Let R be a dyadic rectangle and suppose it has both vert-children and horiz-children. Then the optimal partition of R is either: (i) the trivial partition $\{R\}$, or (ii) the union of optimal partitions of the horiz-kids $\hat{\mathcal{P}}(R^{1,0}) \cup \hat{\mathcal{P}}(R^{1,1})$, or (iii) the union of optimal partitions of the vert-kids $\hat{\mathcal{P}}(R^{2,0}) \cup \hat{\mathcal{P}}(R^{2,1})$. Which of these three cases holds can be determined by holding a "tournament", selecting the winner as the smallest of the three numbers

$$\|y - Ave\{y|R\}\|^2_{\ell^2(R)}, CART(R^{1,0}) + CART(R^{1,1}), CART(R^{2,0}) + CART(R^{2,1}).$$

The exception to this rule is of course at the finest scale: a 1-by-1 rectangle has no children, and so the optimal partition of such an R is just the trivial partition $\{R\}$.

By starting from the next-to-finest scale and applying the inheritance property, we can get the optimal partitions of all 2-by-1 rectangles, and of all 1-by-2 rectangles. By going to the next coarser level and applying inheritance, we can get the optimal partitions of all 4-by-1, of all 2-by-2 and of all 1 by 4 rectangles. And so on. Continuing in a fine-to-coarse or 'bottom-up' fashion, we eventually get to the coarsest level and obtain an optimal partition for $[0,n)^2$.

There are $\approx 2n$ dyadic intervals and hence $\approx 4n^2 = 4N$ dyadic rectangles. Each dyadic rectangle is visited once in the main loop of the algorithm and there are at most a certain constant number C of additions and multiplications per visit. Total work: $\leq C \cdot 4 \cdot N$ flops and $\leq 16 \cdot N$ storage locations. See the appendix for a formal description of the algorithm.

3 Best-Ortho-Basis

We now turn attention away from CART. We recall the standard notation for Haar functions in dimension 1. Let I be a dyadic subinterval of $[0,n)$ and let $\chi_I(i) = |I|^{-1/2} 1_I(i)$. If I contains at least 2 points, set $h_I(i) = (1_{I^{(1)}}(i) - 1_{I^{(0)}}(i))|I|^{-1/2}$, where $I^{(1)}$ is the right half of I and $I^{(0)}$ is the left half of I.

Using these, we can build anisotropic Haar functions in 2-dimensions. Let R be a dyadic rectangle $I_1 \times I_2$; we can form 3 atoms

$$\begin{aligned} \phi^0_R(i_1,i_2) &= \chi_{I_1}(i_1)\chi_{I_2}(i_2), \\ \phi^1_R(i_1,i_2) &= h_{I_1}(i_1)\chi_{I_2}(i_2), \\ \phi^2_R(i_1,i_2) &= \chi_{I_1}(i_1)h_{I_2}(i_2). \end{aligned}$$

These are naturally associated with the rectangle R: ϕ^0_R is, up to scaling, the indicator of R, while ϕ^1_R and ϕ^2_R are associated with horizontal and vertical midpoint splits of R.

Adapting terminology proposed by Mallat and Zhang in a different setting, we call the ϕ_R^s *atoms*, and the collection of *all* such atoms ϕ_R^s indexed by (R, s), makes up a *dictionary* of atoms. This dictionary is overcomplete; it contains $\geq 3n^2 \approx 3N$ atoms, while the span of these elements is of dimension only N.

3.1 Anisotropic Haar Bases

Certain structured subcollections of the elements of $\mathcal{D}$ make up orthogonal bases. These subcollections are in correspondence with *complete recursive partitions*, that is to say, recursive dyadic partitions in which all terminal nodes are 1-by-1 rectangles $[i_1, i_1+1)\times[i_2, i_2+1)$ containing a single point $\mathbf{i} = (i_1, i_2)$.

Given a complete recursive partition $\mathcal{P}^*$, the corresponding ortho basis $\mathcal{B}$ is constructed as follows. Let $NT(\mathcal{P}^*)$ be the collection of all rectangles encountered at non-terminal stages of the recursive partitioning leading to $\mathcal{P}^*$. Let $R \in NT(\mathcal{P}^*)$. As R is nonterminal it will be further subdivided in forming $\mathcal{P}^*$ — i.e. it will be split either horizontally or vertically; let $s(R) = 1$ or 2 according to the splitting variable chosen. Then define $\mathcal{B}$ as the collection of all such $\phi_R^{s(R)}$ and $\chi_{[0,n)^2}$:

$$\mathcal{B}(\mathcal{P}^*) = \{\chi_{[0,n)^2}\} \;\cup\; \{\phi_R^{s(R)}\}_{R\in NT(\mathcal{P})} \tag{3.1}$$

Theorem 3.1 *Let $\mathcal{P}^*$ be a complete recursive dyadic partition of $[0,n)^2$ and let $\mathcal{B}(\mathcal{P}^*)$ be constructed as in (3.1). This is an orthobasis for the N-dimensional vector space of $n \times n$ arrays.*

Proof. Indeed, $\mathcal{B}$ has cardinality N, and the elements of $\mathcal{B}$ are normalized and pairwise orthogonal. The pairwise orthogonality comes from two simple facts. Take any two distinct elements in $\mathcal{B}$; then either they have disjoint support, or the support of one is included in the support of the other. In the first instance, orthogonality is immediate; in the second instance orthogonality follows from two observations: (i) one element of the pair, call it ϕ, is supported in a rectangle on which the other element, φ say, is constant; and (ii) the element ϕ has zero mean, and so is orthogonal to any function which is constant on its support, i.e. to φ. $\square$

Each such Basis $\mathcal{B}$ has a fast transform, produced in a fashion similar to the Haar transform in dimension 1. Indeed the coefficients in such a basis can be computed in terms of block averages and differences of block averages. If $S(R) = \sum_{\mathbf{i}\in R} y(\mathbf{i})$ denotes the sum of values in a rectangle R, then of course

$$\langle y, \chi_R\rangle = S(R)\cdot|R|^{-1/2} \tag{3.2}$$

while if $(R^{1,0}, R^{1,1})$ are horizontal kids of R

$$\langle y, \phi_R^1\rangle = (S(R^{1,1}) - S(R^{1,0}))\cdot|R|^{-1/2} \tag{3.3}$$

and if $(R^{2,0}, R^{2,1})$ are vertical kids of R

$$\langle y, \phi_R^2\rangle = (S(R^{2,1}) - S(R^{2,0}))\cdot|R|^{-1/2}. \tag{3.4}$$

These relations are useful because there is a simple "Pyramid-of-Adders" for calculating all $(S(R) : R \in NT(\mathcal{P}^*))$ in order N time. See the appendix for a formal description of the algorithm.

3.2 Best Basis Algorithm

The collection of *all* anisotropic Haar bases and fast transforms makes for a potentially very useful library. It contains bases associated with partitions which subdivide much more finely in i_1 than in i_2 in some subsets of $[0,n)^2$ and more finely in i_2 than in i_1 in other subsets. There is therefore the possibility of finding bases very well adapted to certain anisotropic problems.

How to choose a "best-adapted" basis? In the general framework set up in the context of cosine packet/wavelet packet bases by Coifman and Wickerhauser (1992), one specifies an additive "entropy" functional of the vector $\theta \in R^N$

$$\mathcal{E}(\theta) = \sum_{i=1}^{N} e(\theta_i) \tag{3.5}$$

where $e(t)$ is a scalar function. Coifman and Wickerhauser's original proposal was $e_{CW}(t) = -t^2 \log(t^2)$, but $e_p(t) = |t|^p$, where $0 < p < 2$ also makes sense, as well as other choices—see below. One uses such a functional to evaluate the quality of a basis; if $\theta(f, \mathcal{B})$ denotes the vector of coefficients of the object f in basis $\mathcal{B}$, then $\mathcal{E}(\theta(f, \mathcal{B}))$ is a measure of the usefulness of a basis for representing f, and the best basis $\mathcal{B}$ in a library $\mathcal{L}$ of ortho bases solves the problem

$$\min_{\mathcal{B} \in \mathcal{L}} \quad \mathcal{E}(\theta(f, \mathcal{B})). \tag{3.6}$$

In the specific case of interest, there are as many bases in the library as there are complete recursive partitions. Elementary arguments show that the number of bases is exponential in N.

While such exponential behavior makes brute force calculation of the optimum in (3.6) practically impossible, judicious application of dynamic programming gives a practical algorithm.

In order to express the key analytic feature of the objective functional, we take a more general point of view, and consider the problem of finding an optimal basis for *just the data falling in the dyadic rectangle* R. Each complete recursive dyadic partition of R, $\mathcal{P}^*(R)$ say, leads to an anisotropic Haar basis, $\mathcal{B}(R)$ say, for the collection of n-by-n arrays supported only in R. Hence we can define the optimization problem

$$(BOB(R)) \qquad \hat{\mathcal{B}}(R) = \underset{\mathcal{B}(R)}{\operatorname{argmin}}\, \tilde{\mathcal{E}}(\theta(y, \mathcal{B}(R))).$$

Here $\theta(y, \mathcal{B}(R))$ refers to the coefficients in an anisotropic basis for $\ell^2(R)$, and $\tilde{\mathcal{E}}(\theta) = \sum_{i=2}^{dim(\theta)} e(\theta_i)$ refers to a relative entropy, which ignores the first coordinate. We let $\hat{\mathcal{P}}^*(R)$ denote the corresponding optimal complete recursive dyadic partition of R.

Solutions to $BOB(R)$ have a key inheritance property. Let R be a dyadic rectangle and suppose it has both vert-children and horiz-children. Then the optimal basis of R is generated by a complete recursive dyadic partition $\hat{\mathcal{P}}^*(R)$ formed in one of two ways. This partition is either: (i) the union of optimal partitions of the horiz-children $\hat{\mathcal{P}}^*(R^{1,0}) \cup \hat{\mathcal{P}}^*(R^{1,1})$, or (iii) the union of optimal partitions of the vert-children $\hat{\mathcal{P}}^*(R^{2,0}) \cup \hat{\mathcal{P}}^*(R^{2,1})$. Which of these two cases holds can be determined by holding a "tournament", selecting the winner as the smallest of the numbers

$$BOB(R^{1,0}) + BOB(R^{1,1}) + e_1, BOB(R^{2,0}) + BOB(R^{2,1}) + e_2;$$

where $e_i = e(\alpha^i_R)$.

The exception to this rule is of course at the finest scale: a 2-by-1 or 1 by 2 rectangle has only one complete recursive partition, and no tournament is necessary to select a "best" one.

By starting from the next-to-finest scale and applying the inheritance property, we can get the optimal partitions of all 4-by-1, of all 2-by-2 and of all 1 by 4 rectangles (omitting again the tournament for 4 by 1 and 1 by 4 rectangles). And so on. Continuing in a fine-to-coarse or 'bottom-up' fashion, we eventually get to the coarsest level and obtain an optimal partition for $[0, n)^2$.

Once again there are $\approx 4n^2 = 4N$ dyadic rectangles. Each dyadic rectangle is visited once in the main part of the algorithm and there are at most a certain constant number C of additions and multiplications per visit. Total work: $\leq C \cdot 4 \cdot N$ flops and $\leq 4N$ storage locations. See the appendix for a formal description of the algorithm.

4 Best Basis De-Noising

CART has to do with removing noise from the data y to produce a reconstruction $\hat{f}$ approximating the noiseless data f. The philosophy of BOB is much less specific: it may be used for many purposes, for example in data compression and for fast numerical analysis [5]. The application determines the choice of entropy, and the use of the expansion in the best basis.

To use best-basis ideas for noise removal, one could apply the proposals of Donoho and Johnstone [11]. Define

$$\mathcal{E}_\lambda(\theta) = \sum_1^N \min(\theta_i^2, \lambda^2\sigma^2) \tag{4.1}$$

and obtain an optimal basis

$$\hat{\mathcal{B}} = \min_{\mathcal{B}\in\mathcal{L}} \mathcal{E}_\lambda(\theta(y, \mathcal{B})). \tag{4.2}$$

Then apply hard thresholding in the selected basis, at threshold level $\lambda\sigma$:

$$\hat{\theta}_i = \theta(y, \mathcal{B})_i \cdot 1_{\{|\theta(y,\mathcal{B})_i| > \lambda\sigma\}}, \quad 1 \leq i \leq N. \tag{4.3}$$

Reconstruct the object $\hat{f}$ having coefficients $\hat{\theta}$ in basis $\mathcal{B}$. This is the de-noised object.

[11] developed results, to be discussed in section 7, showing that with an appropriate choice of λ, the empirical basis chosen by this scheme was *near-ideal.*

In the current setting, where $\mathcal{L}$ is the library of anisotropic Haar bases, (4.2) is amenable to treatment by the fast best-basis algorithm of the last section. So it may be computed in order N time. This de-noising estimate, while possessing certain nice characteristics, lacks one of the attractive features of CART: an interpretation as a spatially-adaptive averaging method. Such a spatially-adaptive method would have the form

$$\hat{f}(\mathbf{i}) = \sum_{R\in\mathcal{P}} \langle y, \chi_R\rangle \chi_R(\mathbf{i})$$

giving a piecewise-constant reconstruction based on rectangular averages of the noisy data y over rectangles R. Here the partition $\mathcal{P} = \mathcal{P}(y)$ would be chosen data-adaptively, and once the partition were chosen, the reconstruction would take a simple form of averaging. While we will mention this procedure further below, and use its properties, we mention it now only to show that threshold de-noising in a Best-Ortho-Basis is not identical to CART.

5 Tree Constraints in the 1-d Haar System

In the context of the ordinary 1-d Haar transform, Joachim Engel (Engel, 1994) has shown that a special type of reconstruction in the Haar system can be related to recursive partitioning. Let, temporarily, $y = (y_i)_{i=0}^{n-1}$ and suppose $y_i = g(i) + v_i$, with v_i noise. Consider reconstructions $\hat{g}$ of the form

$$\hat{g}(i) = \bar{y} + \sum_I w_I \langle y, h_I \rangle \; h_I(i) \tag{5.1}$$

where the sum is over dyadic subintervals of $[0, n)$ and the w_I are scalar "weights". Now impose on the weights (w_I) two constraints

[Tree-i]. *Keep-or-kill.* Each weight is 1 or 0.

[Tree-ii]. *Heredity.* w_I can be 1 only if also $w_{I'} = 1$ whenever $I \subset I'$. If $w_I = 0$ then $w_I = 0$ for every $I' \subset I$.

Each set of weights satisfying these constraints selects the nodes of a dyadic tree T. Engel has called such constraints tree-constraints, and shown that reconstructions obeying these constraints may be put in the form of spatial averages.

Theorem 5.1 *(Engel, 1994) Suppose that $\hat{g}$ defined by (5.1) obeys the tree-constraints (Tree-i) and (Tree-ii). Say that I is terminal if $w_I = 1$ but every interval $I' \subset I$ has $w_{I'} = 0$. The collection of terminal intervals forms a partition $\mathcal{P}$, and*

$$\hat{g}(i) = \sum_{I \in \mathcal{P}} \langle y, \chi_I \rangle \chi_I(i). \tag{5.2}$$

6 Hereditary Constraints and CART

Tree-constraints make sense also in the setting of 2-d anisotropic Haar bases. We consider reconstructions

$$\hat{f}(\mathbf{i}) = \bar{y} + \sum_{R \in NT(\mathcal{P}^*)} w_R \langle y, \phi_R^{s(R)} \rangle \phi_R^{s(R)}(\mathbf{i}) \tag{6.1}$$

where $\mathcal{P}^*$ is a complete recursive dyadic partition, $\{\chi_{[0,n)^2}, (\phi_R^{s(R)})\}$ the associated orthogonal basis, and the weights (w_R) obey two *hereditary constraints*

[Hered-i]. *Keep-or-kill.* Each weight w_R is 0 or 1.

[Hered-ii]. *Heredity.* $w_R = 1$ implies $w_{R'} = 1$ for all ancestors R' of R in $\mathcal{P}^*$; $w_R = 0$ implies $w_{R'} = 0$ for all descendants of R in $\mathcal{P}^*$. We state without proof the analog of Engel's theorem.

Theorem 6.1 *The reconstruction $\hat{f}$ obeying (6.1), [Hered-i], and [Hered-ii] has precisely the form*

$$\hat{f}(\mathbf{i}) = \sum_{R \in \mathcal{P}} \langle y, \chi_R \rangle \chi_R(\mathbf{i})$$

for some possibly incomplete recursive dyadic partition $\mathcal{P}$.

Three questions arise naturally about reconstructions obeying hereditary constraints.

Q1. How to find the best hereditary reconstruction in a given basis?

Q2. How to find the basis in which hereditary reconstruction works best?

Q3. How to efficiently calculate the hereditary best-basis?

All three questions have attractive answers.

6.1 Best Hereditary Reconstruction in Given Basis.

Let T^* denote the complete binary tree of depth $\log_2(N) - 1$. Identifying subtrees $T \subset T^*$ with the weights (w_R) obeying [Hered-i]-[Hered-ii], we write $\hat{f}_{\mathcal{B},T}$ for the reconstruction (6.1) in basis $\mathcal{B}$ having weights (w_I) associated with the tree T.

We define the "best" hereditary reconstruction in terms of the hereditary CPRSS

$$CPRSS(T; \lambda, \mathcal{B}) = \|y - \hat{f}_{\mathcal{B},T}\|_{\ell^2}^2 + \lambda \#(T). \tag{6.2}$$

and the optimization problem is the one achieving the minimal CPRSS among all such reconstructions:

$$\min_{T \subset T^*} CPRSS(T; \lambda, \mathcal{B}). \tag{6.3}$$

By orthogonality of the basis $\mathcal{B}$, we can reformulate this in terms of coordinates. Let $\theta = \theta(y, \mathcal{B})$ denote the vector of coordinates and $(w_R \theta_R(y, \mathcal{B}))$ denote the same vector after applying weights w_R associated with the subtree T. The we have the following equivalent form of (6.2):

$$CPRSS(T) = \sum_R ((w_R - 1)^2 \theta_R^2 + \lambda w_R)$$

This quantity has an inheritance property, which we express as follows. Let $T^*(R)$ denote the complete tree of depth $\log_2(\#R) - 1$ and define the optimization problem

$$(Hered(R)) \qquad \min_{T \subset T^*(R)} \sum_R ((w_R - 1)^2 \theta_R^2 + \lambda w_R)$$

The optimization problem implicitly defines an optimal subtree $\hat{T}(R)$. The inheritance property: the optimal subtree $\hat{T}(R)$ is a function of the optimal subtrees of the children problems $\hat{T}(R^{s(R),b})$, $b = 0, 1$. The tree $\hat{T}(R)$ is either: the empty subtree, or else it has $\hat{T}(R^{s(R),b})$ as subtrees joined at $root(\hat{T}(R))$.

It follows by this inheritance property that the optimal subtree may be computed by a bottom-up pruning exactly as in the optimal-pruning algorithm of CART, Algorithm 10.1, page 294 of the CART book. Hence, a minimizing subtree may be found in order N time. A formal statement of the algorithm is given in the appendix.

6.2 Best Basis for Hereditary Reconstruction

We can define the quality of a basis for hereditary reconstruction by considering the optimum value of the CPRSS functional over all hereditary reconstructions in that basis. Hence, define the Hereditary entropy

$$\mathcal{H}_\lambda(\mathcal{B}) = \min_{T \subset T^*} CPRSS(T; \lambda, \mathcal{B}). \tag{6.4}$$

A Best Basis for hereditary reconstruction is then the solution of

$$\min_{\mathcal{B} \subset \mathcal{L}} \mathcal{H}_\lambda(\mathcal{B}), \tag{6.5}$$

where $\mathcal{L}$ is a library of orthogonal bases. This may be motivated in two ways. First, the goal is intrinsically reasonable, as it seeks a best tradeoff, over all bases and all subtrees, of complexity $\#(T)$ against fidelity to the data $\|y - \hat{f}_{\mathcal{B},T}\|_2^2$. Second, we will prove below that the reconstruction obtained in the optimum basis has a near-ideal mean-squared error.

6.3 Fast Algorithm via CART

The "entropy" $\mathcal{H}_\lambda(\mathcal{B})$ is not an additive functional $\sum_{i=1}^N e(\theta_i(y, \mathcal{B}))$ of the coordinates of y in basis $\mathcal{B}$. Therefore the best-basis algorithm of Section 3, strictly speaking, does not apply. Luckily, we can use the fast CART algorithm. By now this is obvious, we summarize this fact formally, though without writing out the proof.

Theorem 6.2 *When λ is the same in both, CART and BOB with hereditary constraints have the same answers. More precisely,*

$$\min_{\mathcal{B} \in \mathcal{L}} \mathcal{H}_\lambda(\mathcal{B}) = \min_{\mathcal{P}} CPRSS(\mathcal{P}, \lambda) \tag{6.6}$$

The solution of the best-basis problem (6.5) gives, explicitly, an anisotropic basis $\hat{\mathcal{B}}$ and, implicitly by (6.3), an optimal subtree $\hat{T}$; the solution of the CART problem (2.4) gives an optimizing partition $\hat{\mathcal{P}}$, and we have

$$\hat{f}_{\hat{\mathcal{B}},\hat{T}}(\cdot) = \hat{f}(\cdot|\hat{\mathcal{P}}).$$

Remark 1. Although $\mathcal{H}_\lambda(\mathcal{B})$ is not additive, a fast algorithm for computing it is available — the dyadic CART algorithm of Section 1! This shows that fast Best Basis algorithms may exist for certain non-additive entropies.

Remark 2. Although CART and Best-Ortho-Basis are *not* the same in general, *in this case*, with a specific set of definitions of Best Ortho Basis and a specific set of restrictions on the splits employed by CART, the two methods are the same.

7 Oracle Inequalities

CART and BOB define objects which are the solutions of certain optimization problems and hence are in some sense "optimal". However, we should stress that they are optimal *only* in the very artificial sense that they solve certain optimization problems we have defined.

We now turn to the question of *performance* according to externally defined standards, which will lead ultimately to a proof of our main result. This will entail a certain kind of "near-optimality" with a more significant and useful meaning.

In accordance with the philosophy laid out in [9], we approach this from two points of view. First, there is a statistical decision theory component of the problem which we deal with in this section; Second, there is a harmonic analysis component of the problem, which we deal with in the following section.

7.1 Oracle Inequalities

Once more we are in the model (2.1) and we wish to recover f with small mean-squared error. We evaluate an estimator $\hat{f} = \hat{f}(y)$ by its risk

$$R(\hat{f}, f) = E\|\hat{f}(y) - f\|_2^2.$$

Suppose we have a collection of estimators $\hat{\Phi} = \{\hat{f}(\cdot)\}$; we wish to use the one best adapted to the problem at hand. The best performance we can hope for is what Donoho and Johnstone (1994) call the *ideal risk*

$$\mathcal{R}^*(\hat{\Phi}, f) = \inf\{R(\hat{f}, f) : \hat{f} \in \hat{\Phi}\}.$$

We call this ideal because it can be attained only with an oracle, who in full knowledge of the underlying f (but not revealing this to us) selects the best estimator *for this* f from the collection $\hat{\Phi}$.

We optimistically propose $\mathcal{R}^*(\hat{\Phi}, f)$ as a target, and seek true estimators which can approach this target. It turns out that in several examples, one can find estimators which achieve this to within logarithmic terms. The inequalities which establish this are of the form

$$R(\hat{f}^*, f) \leq Const \cdot \log(N) \cdot (\sigma^2 + \mathcal{R}^*(\hat{\Phi}, f)) \quad \forall f$$

which Donoho and Johnstone [10, 11, 12] call *oracle inequalities*, because they compare the risk of valid procedures with the risk achieveable by idealized procedures which depend on oracles.

7.2 Example: Keep-or-Kill De-Noising

Suppose we are operating in a fixed orthogonal basis $\mathcal{B}$ and consider the family $\hat{\Phi}$ of estimators defined by keeping or killing empirical coefficients in the basis $\mathcal{B}$. Such estimators $\hat{f}(y; w)$ are given in the basis $\mathcal{B}$ by

$$\theta_i(\hat{f}, \mathcal{B}) = w_i \cdot \theta_i(y, \mathcal{B}), \quad i = 1, \ldots, N$$

where each weight w_i is either 0 or 1. Such estimators have long been considered in the context of Fourier series estimation, where the basis is the fourier basis, the coefficients are Fourier coefficients, and the w_i are 1 only for $1 \leq i \leq k$ for some frequency cutoff k. Estimators of this form have also been considered by Donoho and Johnstone in the context where $\mathcal{B}$ is a wavelet basis [10]; in that setting the unit weights are ideally chosen at sites of important spatial variability.

Formally then $\hat{\Phi} = \{\hat{f}(\cdot; w) : w \in \{0,1\}^N\}$ is the collection of all keep-or-kill estimators in the fixed basis $\mathcal{B}$. In [10], Donoho and Johnstone studied the nonlinear estimator $\hat{f}^*$, defined in the basis $\mathcal{B}$ by hard thresholding

$$\theta_i(\hat{f}^*(y), \mathcal{B}) = \eta_{\sqrt{\lambda_N}}(\theta_i(y, \mathcal{B})), \quad i = 1, \ldots, N$$

where $\eta_t(y) = 1_{\{|y|>t\}}(y) \cdot \mathrm{sgn}(y)$ is the hard thresholding nonlinearity and $\lambda_N = \sigma^2 2 \log(N)$. They showed that $\hat{f}^*$ obeys the oracle inequality

$$R(\hat{f}^*, f) \leq 2 \cdot \log(N) \cdot (\sigma^2 + \mathcal{R}^*(\hat{\Phi}, f)) \quad \forall f$$

as soon as $N \geq 4$. In short, simple thresholding comes within log-terms of ideal keep-or-kill behavior.

The reader will find it instructive to note that the estimator $\hat{f}^*$ can also be defined as the solution of the optimization problem

$$\min_{w \in \{0,1\}^N} \|y - \hat{f}(y; w)\|^2 + \lambda_N \#\{i : w_i \neq 0\}$$

This is, of course, a complexity-penalized RSS, with penalty term λ_N. Thus the near-ideal estimator is the solution of a minimum CPRSS principle.

7.3 Example: Best-Basis De-Noising

Suppose now we are operating in a library $\mathcal{L}$ of orthogonal bases $\mathcal{B}$ and consider the family $\hat{\Phi}$ of estimators defined by keeping or killing empirical coefficients in *some* basis $\mathcal{B} \in \mathcal{L}$. Such estimators $\hat{f}(y; w, \mathcal{B})$ are of the form

$$\theta_i(\hat{f}, \mathcal{B}) = w_i \cdot \theta_i(y, \mathcal{B}), \quad i = 1, \ldots, N$$

where each weight w_i is either 0 or 1.

Formally $\hat{\Phi} = \{\hat{f}(\cdot; w, \mathcal{B}) : w \in \{0,1\}^N,\ \mathcal{B} \in \mathcal{L}\}$. For obvious reasons, we call $\mathcal{R}^*(\hat{\Phi}, f)$ also $\mathcal{R}^*(\text{IDEAL BASIS}, f)$ In [11], Donoho and Johnstone developed a nonlinear estimator $\hat{f}^*$, with near-ideal properties; it is precisely the best-basis-de-noising estimator defined in section 4; see (4.1)-(4.3). In detail they supposed that among all bases in the library there are at most M distinct elements. They suppose that we pick $\xi > 8$, and set $t = \sqrt{2\log_e(M)}$, then with $\lambda = (\xi \cdot (1+t))^2$, they prove a result almost as strong as the following, which we prove in the appendix below.

Theorem 7.1 *For an appropriate constant $A(\xi)$, the BOB estimator obeys the oracle inequality*

$$R(\hat{f}^*, f) \le A(\xi) \cdot \lambda \cdot (\sigma^2 + \mathcal{R}^*(\hat{\Phi}, f)) \quad \forall f \tag{7.1}$$

In short, empirical best basis (with an appropriate entropy) comes within log-terms of ideal keep-or-kill behavior in an ideal basis. In the specific case of the library of anisotropic Haar bases, $M = 4N$, and so for a fixed choice of ξ, (7.1) becomes

$$R(\hat{f}^*, f) \le Const \cdot \log(N) \cdot (\sigma^2 + \mathcal{R}^*(\text{IDEAL BASIS}, f)) \quad \forall f.$$

7.4 Example: CART

Oracle inequalities for CART are now easy to state. Suppose now we are operating in the library $\mathcal{L}$ of anisotropic Haar bases and consider the family $\hat{\Phi}_{Tree}$ of *hereditary linear estimators*, i.e. estimators defined by keeping or killing the empirical coefficients in some basis $\mathcal{B} \in \mathcal{L}$, where the coefficients that are kept fall in a tree pattern T. Such estimators $\hat{f}(y; T, \mathcal{B})$ are of the form

$$\theta_i(\hat{f}, \mathcal{B}) = w_i \cdot \theta_i(y, \mathcal{B}), \quad i = 1, \ldots, N$$

where each weight w_i is either 0 or 1, and the nonzero w form a tree.

Formally $\hat{\Phi}_{Tree} = \{\hat{f}(\cdot; T, \mathcal{B}) : T \subset T^*\ \mathcal{B} \in \mathcal{L}\}$ be the collection of all hereditary linear estimators in any anisotropic Haar basis. The ideal risk $\mathcal{R}^*(\hat{\Phi}_{Tree}, f)$ is just the risk of CART applied in an ideal partition selected by an oracle $\mathcal{P}$. So call this $\mathcal{R}^*(\text{IDEAL CART}, f)$

Consider now the dyadic CART estimator $\hat{f}^*$ defined with λ exactly as specified in the best-basis de-noising setting of the last subsection. So for $\xi > 8$, set $\lambda = (\xi \cdot (1 + \sqrt{2\log_e(4N)}))^2$. We prove the following in the appendix.

Theorem 7.2 *For all $N \ge 1$, the dyadic CART estimator obeys the oracle inequality*

$$R(\hat{f}^{*,\lambda}, f) \le Const \cdot \log(N) \cdot (\sigma^2 + \mathcal{R}^*(\text{IDEAL CART}, f)) \quad \forall f$$

In short, empirical dyadic CART (with an appropriate entropy) comes within log-terms of ideal dyadic CART.

8 Mixed Smoothness Spaces

We now change gears slightly, and consider harmonic analysis questions. Specifically we are going to show that anisotropic Haar bases are particularly well-adapted to dealing with classes of functions having mixed smoothness.

We denote now by f a function $f(x,y)$ defined on $[0,1]^2$, rather than an array of pixel values. We consider objects of different smoothnesses in different directions; the specific notion of smoothness we use is based on what Temlyakov calls Nikol'skii classes [17]. Define the finite difference operators $(D^1_h f)(x,y) = f(x+h,y) - f(x,y)$ and $(D^2_h f)(x,y) = f(x,y+h) - f(x,y)$. For δ_1, δ_2 satisfying $0 \le \delta_i \le 1$ define the mixed smoothness class

$$\begin{aligned}\mathcal{F}^{\delta_1,\delta_2}_p(C) = \{f : \|f\|_p \le C, \quad & \|D^1_h f\|_{L^p(Q^1_h)} \le C h^{\delta_1} \; h \in (0,1), \\ & \|D^2_h f\|_{L^p(Q^2_h)} \le C h^{\delta_2} \; h \in (0,1)\}\end{aligned}$$

where $Q^1_h = [0, 1-h) \times [0,1]$ and $Q^2_h = [0,1] \times [0, 1-h)$. This contains objects of genuinely mixed smoothness whenever $\delta_1 \neq \delta_2$. The usual smoothness spaces (Hölder, Sobolev, Triebel, $\cdots$) involve equal degrees of smoothness in differrent directions, are are sometimes called "isotropic," so that classes like $\mathcal{F}^{\delta_1,\delta_2}_p(C)$ would be called "anisotropic."

8.1 Spatially Uniform Anisotropic Bases

Definition 8.1 *A* sequential partitioning of j into two parts *is a pair of series of integers* $j_1(j), j_2(j),\ j = 0, 1, 2, \cdots$ *obeying*

- *Initialization.* $j_1(0) = j_2(0) = 0$.
- *Partition.* $j_1(j) + j_2(j) = j$.
- *Sequential Allocation.*

$$\begin{aligned} j_1(j) &= j_1(j-1) + b_1(j), \quad b_1 \in \{0,1\}, \\ j_2(j) &= j_2(j-1) + b_2(j), \quad b_2 = 1 - b_1. \end{aligned}$$

We can think of two boxes and a sequential scheme where at each stage we put a ball in one of the two boxes. $j_i(j)$ represents the number of balls in box i at stage j, and $b_1 = 1 - b_2$ represents the constraint that only one ball is put into the boxes at each stage.

Definition 8.2 *Consider a sequential partition of j into two parts. The* spatially uniform alternating partition *subordinate to this partition – $SUAP(j_1, j_2)$ – is a complete dyadic recursive partiton formed in a homogeneous fashion: at Stage 1, the square $[0,1]^2$ is split horizontally if $b_1(1) = 1$, and vertically if $b_2(1) = 1$; at Stage 2, each of the two resulting rectangles is split in two, horizontally if $b_1(2) = 1$, vertically if $b_2(2) = 1$; and at Stage j, each of the 2^{j-1} rectangles of volume 2^{-j+1} formed at the previous stage is split vertically if $b_1(j) = 1$, horizontally if $b_2(j) = 1$.*

The recursive partition SUAP(j_1, j_2) defines a series of collections of rectangles as follows. $\mathcal{R}(0)$ consist of the root rectangle, $\mathcal{R}(1)$ consists of the two children of the root, $\mathcal{R}(2)$ of the four children of the rectangles in $R(1)$, etc. In general $\mathcal{R}(j)$ consists of 2^j rectangles of area 2^{-j} each.

This sequence of rectangles defines an orthogonal basis of $L^2([0,1]^2)$ in a fashion similar to the discrete case, with fairly obvious changes due to the change in setting. Let now I denote a dyadic subinterval of $[0,1]$ and $\tilde{\chi}_I(x)$ be the "same function" as $\chi_I(i)$, under the correspondence $x_i \leftrightarrow i/n$ and under the different choice of normalizing measure: $\tilde{\chi}_I(x) = 1_I(x) \cdot \ell(I)^{-1/2}$ where $\ell(I)$ denotes the length of I. Similarly, let $\tilde{h}_I(x) = (1_{I^1}(x) - 1_{I^0}(x)) \cdot \ell(I)^{-1/2}$. Then set $\varphi_R^1 = \tilde{h}_{I_1}(x)\tilde{\chi}_{I_2}(y)$, $\varphi_R^2 = \tilde{\chi}_{I_1}(x)\tilde{h}_{I_2}(y)$. Then set

$$\begin{aligned} \xi_0 &= \chi_{[0,1]^2} \quad , \quad \xi_{0,0} = \varphi_{[0,1]^2}^{b_2(1)}, \\ \xi_{1,R} &= \varphi_R^{b_2(2)} \quad \text{for } R \in \mathcal{R}(1), \\ \xi_{2,R} &= \varphi_R^{b_2(3)} \quad \text{for } R \in \mathcal{R}(2), \end{aligned}$$

and in general

$$\xi_{j,R} = \varphi_R^{b_2(j)} \quad \text{for } R \in \mathcal{R}(j); \tag{8.1}$$

call this the *spatially homogeneous anisotropic basis* SHAB (j_1, j_2).

The coefficients of f in this basis are

$$\bar{f} = Ave_{[0,1]^2}, \qquad \alpha_R = \langle \xi_{j,R}, f \rangle, \quad R \in \mathcal{R}(j). \tag{8.2}$$

8.2 Best Basis for a functional class

Donoho (1993) described a notion of "best-orthogonal-basis" for a functional class $\mathcal{F}$ which describes the kinds of bases in which certain kinds of de-noising and data compression can take place. For this notion, the best basis for a functional class $\mathcal{F}$ is that basis in which the coefficients of members of $\mathcal{F}$ decay fastest.

For a vector θ in sequence space let $|\theta|_{(i)}$ denote the rearranged magnitudes of the coefficients, sorted in decreasing order $|\theta|_{(1)} \geq |\theta|_{(2)} \geq \ldots$. The weak ℓ^τ norm measures the decay of these by

$$\|\theta\|_{wl^\tau} = \sup_{k \geq 1} k^{1/\tau} |\theta|_{(k)}.$$

This measures decay of the coefficients since $\|\theta\|_{wl^\tau} \leq C$ implies $|\theta|_{(k)} \leq Ck^{-1/\tau}$, $k = 1, 2, \ldots$.

Now, with $\theta = (\theta_i(f, \mathcal{B}))$ the coefficients of f in an orthogonal basis $\mathcal{B}$, a given functional class $\mathcal{F}$ maps to a coefficient body $\Theta(\mathcal{F}, \mathcal{B}) = \{(\theta_i(f, \mathcal{B}))_i : f \in \mathcal{F}\}$. For such a set Θ, we say $\Theta \subset wl^\tau$ if

$$\sup\{\|\theta\|_{wl^\tau} : \theta \in \Theta\} < \infty.$$

Definition 8.3 *We call the* critical exponent *of* Θ *the number* $\tau^*(\Theta)$ *obtained as the infimum of all* τ *for which* $\Theta \subset wl^\tau$.

Assuming that $\mathcal{F}$ is a subset of L^2, $0 \leq \tau^*(\Theta) \leq 2$.

From this point of view a "best basis" for $\mathcal{F}$ is any basis $\mathcal{B}^*$ which minimizes the critical exponent:

$$\tau^*(\Theta(\mathcal{F}, \mathcal{B}^*)) = \min_{\mathcal{B}} \tau^*(\Theta(\mathcal{F}, \mathcal{B})). \tag{8.3}$$

In such a basis the rearranged coefficients will be the most rapidly decaying, among all ortho bases.

8.3 Best Anisotropic Bases

With this background, it is interesting to ask about the decay properties of coefficients in different spatially homogeneous anisotropic bases. We might hope to identify a basis $\mathcal{B}$ within the class of anisotropic Haar bases satisfying (8.3) among all bases. In fact we can. The key fact is this upper bound.

Lemma 8.4 *Let $f \in \mathcal{F}_p^{\delta_1,\delta_2}(C)$. If $b_1(j) = 1$,*

$$\Big(\sum_{R \in \mathcal{R}(j)} |\alpha_R|^p \Big)^{1/p} \leq C \cdot 2^{-j_1 \delta_1} \cdot (2^{-j})^{1/2-1/p}, \tag{8.4}$$

while if $b_2(j) = 1$

$$\Big(\sum_{R \in \mathcal{R}(j)} |\alpha_R|^p \Big)^{1/p} \leq C \cdot 2^{-j_2 \delta_2} \cdot (2^{-j})^{1/2-1/p}. \tag{8.5}$$

Now a choice of spatially uniform anisotropic partition which would make optimal use of these expressions as a function of j would arrange things so that the decrease of the largest of the two expressions went fastest in j. Thus optimal use of Lemma 8.4 leads to the problem of constructing a sequential partition of j into parts that optimizes the rate of decay of

$$\max(2^{-j_1(j)\delta_1}, 2^{-j_2(j)\delta_2}) \tag{8.6}$$

as a function of j.

There is an obvious limit on how well this can be done. Consider optimizing (8.6) subject to only the constraints $j_1(j) + j_2(j) = j$ and $j_i \geq 0$, i.e. without imposing the requirement that the j_i be integers, or be sequentially chosen. The solution is $j_1(j) = \delta_2/(\delta_1 + \delta_2) \cdot j$ and $j_2(j) = \delta_1/(\delta_1 + \delta_2) \cdot j$, achieving an optimally small value of

$$2^{-j\rho}, \qquad \rho = \frac{\delta_1 \delta_2}{\delta_1 + \delta_2}, \tag{8.7}$$

in (8.6). We cannot hope to do better than this, once we re-impose the constraints associated with a sequential partition. But we can come close.

Definition 8.5 *For a given pair of "exponents" δ_1, δ_2 obeying $0 < \delta_i \leq 1$, we call an* optimal sequential partition of j *a sequential partitioning (j_1^*, j_2^*) obtained as follows:*

1. *Start from $j_1^*(0) = j_1^*(0) = 0$.*
2. *At stage $j+1$, allocate $b_1(j)$ and $b_2(j)$ as follows.*

   a. *If $j_1^*(j-1)\delta_1 = j_2^*(j-1)\delta_2$ "allocate the ball to whichever box has the smaller exponent": $b_1(j) = 1$ if $\delta_1 < \delta_2$.*

   b. *If $j_1^*(j-1)\delta_1 \neq j_2^*(j-1)\delta_2$ "allocate the ball to whichever box has the smaller product": $j_i^*(j-1)\delta_i$.*

This so-called optimal sequential partitioning of j is a greedy stepwise minimization of objective (8.6) It turns out that it is near-optimal, even among nonsequential partitions.

Lemma 8.6 *For $0 < \delta_1, \delta_2 \le 1$, and $\rho = \frac{\delta_1 \delta_2}{\delta_1 + \delta_2}$,*

$$\max(2^{-j_1^*(j)\delta_1},\ 2^{-j_2^*(j)\delta_2}) \le 2 \cdot 2^{-j\rho}. \tag{8.8}$$

Due to (8.7), this is essentially optimal within the class of sequential partitions of j. Inspired by this, we propose the following

Definition 8.7 *We call Best Anisotropic Basis $BAB(\delta_1, \delta_2)$ the anisotropic basis of $L^2[0,1]^2$ defined using $SHAB(j_1^*, j_2^*)$.*

Combining this Lemma with Lemma 8.4 above

Corollary 8.8 *If we use the $BAB(\delta_1, \delta_2)$ then for $\rho = \frac{\delta_1 \delta_2}{\delta_1 + \delta_2}$,*

$$\Big(\sum_{R \in \mathcal{R}(j)} |\alpha_R|^p \Big)^{1/p} \le 2 \cdot C \cdot 2^{-j(\rho + 1/2 - 1/p)}. \tag{8.9}$$

8.4 Optimality of BAB

Armed with Corollary 8.8, it is possible to justify Definition 8.7 and prove that $BAB(\delta_1, \delta_2)$ is an optimal basis in the sense of section 8.2.

Theorem 8.9 *Let $\mathcal{L}$ denote the collection of all orthogonal bases for $L^2[0,1]^2$.*

$$\tau^*(\Theta(\mathcal{F}_p^{\delta_1,\delta_2}(C), BAB(\delta_1, \delta_2))) = \min_{\mathcal{B} \in \mathcal{L}} \tau^*(\Theta(\mathcal{F}_p^{\delta_1,\delta_2}(C), \mathcal{B})). \tag{8.10}$$

The proof is a consequence of three lemmas, all of which are proved in the appendix. The first gives an evaluation of the critical exponent for $BAB(\delta_1, \delta_2)$.

Lemma 8.10

$$\tau^*(\Theta(\mathcal{F}_p^{\delta_1,\delta_2}(C), BAB(\delta_1, \delta_2))) = 2/(2\rho + 1). \tag{8.11}$$

The optimality of this exponent follows from a lower bound technique developed at greater length in [8]. First, a definition: an *orthogonal hypercube* $\mathcal{H}$ of dimension m and side ϵ is a collection of all sums $g_0 + \sum_{i=1}^m \alpha_i g_i$ where the g_i are orthonormal functions and the $|\alpha_i| \le \epsilon$.

Lemma 8.11 *Suppose $\mathcal{F}$ contains a sequence $\mathcal{H}_j$ of orthogonal hypercubes of dimension m_j and side ϵ_j where $\epsilon_j \to 0$, $m_j \to \infty$,*

$$m_j^{1/\tau} \epsilon_j \ge c_0 > 0$$

Let $\mathcal{L}$ denote any collection of orthogonal bases.

$$\inf_{\mathcal{B} \in \mathcal{L}} \tau^*(\Theta(\mathcal{F}, \mathcal{B})) \ge \tau.$$

Lemma 8.12 *Each class $\mathcal{F}_p^{\delta_1,\delta_2}(C)$ contains a sequence $\mathcal{H}_j$ of orthogonal hypercubes of dimension $m_j = 2^j$ and side ϵ_j where $\epsilon_j \to 0$, $m_j \to \infty$,*

$$m_j^{1/\tau} \epsilon_j \ge K \cdot C,$$

with K a fixed constant, and $\tau = 2/(2\rho + 1)$.

9 Near-Minimaxity of BOB

As a result of the harmonic analysis in section 8 and the ideas in [8], we know that $BAB(\delta_1, \delta_2)$ is the best basis in which to apply ideal keep-or-kill estimates. This is the key stepping stone to our main result.

In this section we show that the risk for ideal keep-or-kill in $BAB(\delta_1, \delta_2)$ is within constants of the minimax risk over each $\mathcal{F}_p^{\delta_1,\delta_2}(C)$. From the oracle inequality of section 7.2, we know that empirical basis selection, as in (4.1)-(4.3), which empirically selects a basis and applies thresholding within it, will always be nearly as good as ideal keep-or-kill in $BAB(\delta_1, \delta_2)$ – even though it makes no assumptions on δ_1 or δ_2. This means that empirical best-basis de-noising obeys a near-minimaxity result like Theorem 1.1.

Theorem 9.1 *Best-Basis De-Noising, defined in section 4 above, with λ defined as in section 7.2 above, comes within logarithmic factors of minimax over each functional class $\mathcal{F}_p^{\delta_1,\delta_2}(C)$, $0 < \delta_1, \delta_2 \le 1$, $C > 0$, $p \in (0, \infty)$. If $\hat{f}^{*,\lambda}$ denotes the best-basis-denoising estimator*

$$\sup_{\mathcal{F}} E\|\hat{f}^{*,\lambda} - \bar{f}\|_{\ell^2}^2 \le \mathit{Const}(\delta_1, \delta_2, p) \cdot \log(n) \cdot M^*(\sigma, n; \mathcal{F}) \qquad \text{as } \; n \to \infty. \tag{9.1}$$

for each $\mathcal{F} \in \mathcal{MS}$.

The key arguments to prove Theorem 9.1 are given in sections 9.1 and 9.4 below. Our main result – Theorem 1.1 – will be proved in section 10 by using some of those results a second time.

9.1 Lower Bound on the Minimax Risk

We first study the minimax risk and show that it obeys the lower bound

$$M^*(\sigma, n; \mathcal{F}_p^{\delta_1,\delta_2}(C)) \ge K(\delta_1, \delta_2) \cdot (C^2)^{1-r} (\epsilon^2)^r, \quad \text{as} \qquad \epsilon = \sigma/\sqrt{N} \to 0. \tag{9.2}$$

where $r = \frac{2\rho}{2\rho+1}$.

We use the method of cubical subproblems. Modified definition: in this section, by *orthogonal hypercube* $\mathcal{H}$ of dimension m and side δ we mean a collection of all sums $g_0 + \sum_{k=1}^m \alpha_k g_k$ where the $g_k = g_k(i_1, i_2)$ are n by n arrays, orthonormal with respect to the specially normalized ℓ_N^2 norm

$$\frac{1}{\sqrt{N}} \sum_{i_1,i_2} g_k(i_1, i_2) g_{k'}(i_1, i_2) = 1_{\{k=k'\}},$$

and all the $|\alpha_i| \le \delta$. The following Lemma may be proved as in [12].

Lemma 9.2 *Let $\epsilon = \sigma/\sqrt{N}$. Suppose a class $\mathcal{F}$ contains an orthogonal hypercube of side-length $\epsilon \le \delta < \frac{11}{10}\epsilon$ and dimension $m(\epsilon)$. Then, for an absolute constant $A > 1/10$,*

$$M^*(\sigma, n; \mathcal{F}) \;\ge\; A \cdot m(\epsilon) \cdot \epsilon^2. \tag{9.3}$$

To make effective use of this, we seek cubes of sufficiently high dimension and prescribed sidelength. The following lemma is proved in the Appendix.

Lemma 9.3 *Let $\epsilon = \sigma/\sqrt{N}$. Each class $\mathcal{F}_p^{\delta_1,\delta_2}(C)$ contains orthogonal hypercubes (orthogonal with respect to ℓ_N^2 norm) of sidelength $\delta = \epsilon \cdot (1 + o(1))$ and dimension $m(\epsilon, C)$ where*

$$m(\epsilon, C) \geq K(\delta_1, \delta_2) \cdot (C/\epsilon)^{\frac{2}{2\rho+1}}, \quad 0 < \epsilon < \epsilon_0, \tag{9.4}$$

and

$$\rho = \frac{\delta_1 \cdot \delta_2}{\delta_1 + \delta_2}. \tag{9.5}$$

Combining these two lemmas gives the lower bound (9.2).

9.2 Equivalent Estimation Problems

Sections 2-7 of this paper work in a setting of n by n arrays. Section 8 works in a setting of functions on the continuum unit square. Theorem 9.1 is based on a combination of both points of view.

From the viewpoint of Sections 2-7 one would naturally consider applying CART and BOB estimators to data $y_{\mathbf{i}}$, $\mathbf{i} \in [0, n)^2$. Suppose instead that we define the rescaled data

$$\tilde{y}_{\mathbf{i}} = N^{-1/2} y_{\mathbf{i}}, \quad \mathbf{i} \in [0, n)^2,$$

and also define $\epsilon = \sigma/n = \sigma/\sqrt{N}$. The results we get in applying (appropriately calibrated) CART or BOB to such data are (obviously) proportional to the results we get in applying the same techniques to the unscaled data.

There is a connection between these rescaled data and data about the function f on the continuum square. Let R denote both a dyadic rectangle of $[0, n)^2$ and the "same" rectangle on the continuum square $[0, 1]^2$. Recall that $\varphi_R^1(x, y)$ denotes a function on the continuum square $[0, 1]^2$ normalized to $L^2[0, 1]^2$-norm 1, and $\phi_R^1(i_1, i_2) = h_{I_1}(i_1)\chi_{I_2}(i_2)$ is the "same" function, only on the grid $[0, n)^2$ and normalized to $\ell^2(N)$ norm 1. Then

$$\langle \tilde{y}_{\mathbf{i}}, \phi_R^1 \rangle_{\ell^2(N)} = \langle f, \varphi_R^1 \rangle_{L^2[0,1]^2} + \epsilon \cdot z_R^1,$$

where the z_R^1 are $N(0, 1)$, and independent in rectangles which are disjoint. Similar relationships hold between ϕ_R^2 and φ_R^2.

Hence the discrete-basis analysis of rescaled data $\tilde{y}_{\mathbf{i}}$ has the interpretation of giving noisy measurements about the continuum coefficients of f – and vice versa. Moreover, suppose that $\mathcal{P}^*$ is a complete dyadic recursive partition of the discrete grid $[0, n)^2$ and we consider only the coefficients attached to rectangles in the nonterminal nodes of this partition. The partial reconstruction of f from just those coefficients is simply the collection of f's pixel level averages; formally, if we put

$$f(i_1, i_2) = \sum_{R \in NT(\mathcal{P}^*)} \alpha_R \phi_R^{s(R)}$$

and

$$\bar{f}(x, y) = \sum_{R \in NT(\mathcal{P}^*)} \alpha_R \varphi_R^{s(R)}(x, y)$$

then $\bar{f}(x, y)$ takes the value $f(i_1, i_2)$ throughout the rectangle $[i_1/n, (i_1+1)/n) \times [i_2/n, (i_2+1)/n)$.

Consider the problem of estimating $(\alpha_R^s(R))_{R\in NT(\mathcal{P}^*)}$ from noisy data $\langle f, \varphi_R^{s(R)}\rangle_{L^2[0,1]^2} + \epsilon \cdot z_R^{s(R)}$. By Parseval, the squared ℓ^2 risk

$$\epsilon^2 + E \sum_{R\in NT(\mathcal{P}^*)} (\hat{\alpha}_R^{s(R)} - \alpha_R^{s(R)})^2 = N^{-1} \cdot E\|\hat{f} - \bar{f}\|^2_{\ell^2(N)} \tag{9.6}$$

and so the mean-squared error in the coefficient domain gives us the mean-squared error for recovery of pixel-level averages in the other domain.

9.3 Discrete and Continuous Partitionings

Consider now $BAB(\delta_1, \delta_2)$ for a given δ_1, δ_2 pair. This corresponds to an infinite sequence of families $\mathcal{R}(j)$, each family partitioning the continuum square $[0,1]^2$ by congruent rectangles of area 2^{-j}.

Such a sequence of partitions usually does not have, for its initial segment $\log_2(N)$ members, a sequence of partitions which can be interpreted as a sequence of partitions for the discrete square $0 \le i_1, i_2 < n$. A sequence of partitions for the discrete square also has the constraint that out of the first $\log_2(N)$ splits, exactly half will be vertical and half horizontal. Put another way, if we consider some BAB, those rectangles which are not too narrow in any direction – i.e. where each sidelength exceeds $1/n$ – also correspond to rectangles in a complete dyadic recursive partition of the discrete square $[0,n)^2$. But there exist BAB (for example those with $\min(\delta_1, \delta_2)$ close to zero and $\max(\delta_1, \delta_2)$ close to one) which, at some level j between $\log_2(n)$ and $\log_2(N)$, have already split in a certain direction more than $\log_2(n)$ times. Consequently, the continuum BAB is not quite available in the analysis of finite datasets.

On the other hand, in the analysis of finite datasets, there are available bases which achieve the same estimates of coefficient decay as in the continuum case.

Definition 9.4 *For a given pair of exponents (δ_1, δ_2), and whole number J, we call a Balanced Finite Optimal Sequential Partition, an application of the optimal sequential partitioning rule of definition 8.5, with two extra rules:*

3. *The process stops at stage $2J$. There are at most $2J$ "balls".*
4. *The process must preserve $j_i^*(j) \le J$. Once a certain "box" has "J balls", all remaining allocations of "balls" are to the "other box".*

Lemma 9.5 $0 < \delta_1, \delta_2 \le 1$, *let $j_i^*(j)$ denote the result of a Balanced Finite Optimal Sequential Partitioning. Let $\mathcal{R}^*(j)$ denote the associated collection of rectangles. With $\rho = \frac{\delta_1\delta_2}{\delta_1+\delta_2}$,*

$$\Big(\sum_{R\in\mathcal{R}^*(j)} |\alpha_R|^p\Big)^{1/p} \le 2 \cdot C \cdot 2^{-j(\rho+1/2-1/p)}. \tag{9.7}$$

The proof is simply to inspect the proof of Corollary 8.7 and notice that the constraint preventing allocation of "balls" to certain "boxes" means that in certain expressions one can replace terms like

$$\max(2^{-j_1^*(j)\delta_1},\ 2^{-j_2^*(j)\delta_2})$$

by the even smaller

$$\min(2^{-j_1^*(j)\delta_1},\ 2^{-j_2^*(j)\delta_2}).$$

9.4 Upper Bound on Ideal Risk

We now study the ideal risk and show that it obeys an upper bound similar in form to the lower bound of section 9.1. Starting now, let $BAB^*(\delta_1, \delta_2)$ denote the modified basis described in the previous subsection.

Lemma 9.6 *Let $\mathcal{R}(\text{Keep-Kill}, f; \epsilon)$ be the ideal risk for keep-kill estimation in $BAB^*(\delta_1, \delta_2)$. Then with $r = \frac{2\rho}{2\rho+1}$,*

$$\sup_{f \in \mathcal{F}_p^{\delta_1,\delta_2}(C)} \mathcal{R}(\text{Keep-Kill}, f; \epsilon) \leq B(\delta_1, \delta_2, p) \cdot (C^2)^{1-r} (\epsilon^2)^r, \quad 0 < \epsilon < \epsilon_0. \tag{9.8}$$

Proof. As in [13] consider the optimization problem

$$m_j(\epsilon; \gamma) = \max \|\theta\|_{\ell^2}^2 \text{ subject to } \|\theta\|_{\ell^\infty} \leq \epsilon \quad \|\theta\|_{\ell^p} \leq \gamma, \quad \theta \in R^{2^j}.$$

By Parseval (9.6) the best possible risk for a purely keep-kill estimate is $\epsilon^2 + \sum \min(\alpha_R^2, \epsilon^2)$. Also, by Lemma 9.5, there are constants $\gamma_j = \gamma_j(C)$ so that for $f \in \mathcal{F}_p^{\delta_1,\delta_2}(C)$,

$$\Big(\sum_{\mathcal{R}^*(j)} |\alpha_R|^p\Big)^{1/p} \leq \gamma_j$$

The largest risk of ideal keep-kill is thus

$$\begin{aligned}
\max_{f \in \mathcal{F}_p^{\delta_1,\delta_2}(C)} \quad & \textstyle\sum_j \sum_{R \in \mathcal{R}^*(j)} \min(\alpha_R^2, \epsilon^2) \\
\leq \quad & \max \sum_j \sum_{\mathcal{R}^*(j)} \min(\alpha_R^2, \epsilon^2) \text{ subj. to } \Big(\sum_{\mathcal{R}^*(j)} |\alpha_R|^p\Big)^{\frac{1}{p}} \leq \gamma_j \\
= \quad & \sum_j m_j(\epsilon, \gamma_j)
\end{aligned}$$

Now [13] gives the explicit evaluation

$$m_j(\epsilon; \gamma) = \min(2^j \epsilon^2, \gamma^p \epsilon^{2-p}, \gamma^2). \tag{9.9}$$

and applying this, we have

$$\sum_j m_j(\epsilon, \gamma_j) \leq (C^2)^{1-r} (\epsilon^2)^r \cdot K(\delta_1, \delta_2, p). \Box \tag{9.10}$$

This is the risk of an ideal de-noising by a keep-or-kill estimator not obeying hereditary constraints.

9.5 Near-Minimaxity of Best-Basis De-Noising

We have so far shown that the ideal risk is within constant factors of the minimax risk. Invoking now the oracle inequality of Theorem 7.1, the worst-case risk of the BOB estimator $\hat{f}$ does not exceed the ideal risk – and hence the minimax risk – by more than a logarithmic factor. This completes the proof of Theorem 9.1.

10 Near-Minimaxity of CART

We now are in a position to complete the proof of Theorem 1.1. We do this by showing that ideal dyadic CART is essentially as good as ideal best-basis de-noising.

Lemma 10.1 *Let $\mathcal{R}(\text{Keep-Kill}, f; \epsilon)$ be the ideal risk for keep-kill estimation in $BAB^*(\delta_1, \delta_2)$. Let $\mathcal{R}(\text{Hered}, f; \epsilon)$ be the ideal risk for hereditary estimation in $BAB^*(\delta_1, \delta_2)$.*

$$\sup_{f \in \mathcal{F}_p^{\delta_1,\delta_2}(C)} \mathcal{R}(\text{Hered}, f; \epsilon) \le B(\delta_1, \delta_2, p) \cdot \sup_{f \in \mathcal{F}_p^{\delta_1,\delta_2}(C)} \mathcal{R}(\text{Keep-Kill}, f; \epsilon). \qquad (10.1)$$

Once this lemma is established, it follows from sections 9.1 and 9.4 that the risk of ideal dyadic CART is within constant factors of the minimax risk. Now the oracle inequality for dyadic CART, Theorem 7.2, shows that the performance of empirical dyadic CART comes within logarithmic factors of the ideal risk for dyadic CART. Theorem 1.1 therefore follows as soon as Lemma 10.1 is established.

To prove the Lemma, note that the ideal keep-or-kill estimator for a function f has nonzero coefficients at sites

$$\mathcal{S}(f) = \{(j, R) : |\alpha_{j,R}(f)| \ge \epsilon\}.$$

This can be modified to a hereditary linear estimator by replacing $\mathcal{S}$ by $\mathcal{S}^*$, the *hereditary cover* of $\mathcal{S}$.

The ideal risk of the hereditary linear estimator $\hat{\alpha}[\mathcal{S}^*]$ obeys

$$\begin{aligned} E\|\hat{\alpha}[\mathcal{S}^*] - \alpha\|_2^2 &= \sum_{(j,R)\notin\mathcal{S}^*} \alpha_{j,R}^2 + \epsilon^2(\#(\mathcal{S}^*) + 1) \\ &\le \sum_{(j,R)\notin\mathcal{S}} \alpha_{j,R}^2 + \epsilon^2(\#(\mathcal{S}^*) + 1) \qquad (\text{as } \mathcal{S} \subset \mathcal{S}^*). \end{aligned}$$

Suppose we could bound $\#\mathcal{S}^* \le A \cdot (\#\mathcal{S})$ for some constant $A > 0$. Then we would have

$$\begin{aligned} E\|\hat{\alpha}[\mathcal{S}^*] - \alpha\|_2^2 &\le \sum_{(j,R)\notin\mathcal{S}} \alpha_{j,R}^2 + \epsilon^2(A\#(\mathcal{S}) + 1) \qquad (\text{as } \#\mathcal{S}^* \le A\#\mathcal{S}) \\ &\le (A+1) \cdot \left(\sum_{(j,R)\notin\mathcal{S}} \alpha_{j,R}^2 + \epsilon^2 \right) \\ &= (A+1) \cdot E\|\hat{\alpha}[\mathcal{S}] - \alpha\|_2^2. \end{aligned}$$

It would then follow that risk bounds derived for keep-or-kill estimators would give rise to proportional risk bounds for hereditary estimators.

While the relation $\#\mathcal{S}^* \le A \cdot (\#\mathcal{S})$ does not hold for every f, a weaker inequality of the same form holds, where one compares the largest possible size of $\#\mathcal{S}(f)$ for an $f \in \mathcal{F}_p^{\delta_1,\delta_2}(C)$ with the largest possible size of $\#\mathcal{S}^*$. The lemma just below establishes this inequality; retracing the logic of the last few displays shows that it immediately implies Lemma 10.1, with $B = A + 1$.

Lemma 10.2 *Define*

$$N(\delta_1, \delta_2, p, C) = \sup\{\#\mathcal{S}(f) : f \in \mathcal{F}_p^{\delta_1,\delta_2}(C)\} \qquad (10.2)$$

the largest number of coefficients used by an ideal keep-kill estimator in treating functions from $\mathcal{F}_p^{\delta_1,\delta_2}(C)(\delta_1,\delta_2)$*. Similarly, let*

$$N^*(\delta_1,\delta_2,p,C) = \sup\{\#\mathcal{S}^*(f) : f \in \mathcal{F}_p^{\delta_1,\delta_2}(C)\} \tag{10.3}$$

be the size of the largest corresponding hereditary cover. Then for a finite positive constant $A = A(\delta_1,\delta_2,p)$*,*

$$N^* \le A(\delta_1,\delta_2,p)\cdot N.$$

Proof. If $\theta = (\theta_i)_{i=1}^d$ is a vector of dimension d satisfying $\|\theta\|_{\ell^p} \le \gamma$, then $\#\{i : |\theta_i| \ge \epsilon\}\epsilon^p \le \gamma^p$ so

$$\#\{i : |\theta_i| \ge \epsilon\} \le (\gamma/\epsilon)^p. \tag{10.4}$$

and of course

$$\#\{i : |\theta_i| \ge \epsilon\} \le d. \tag{10.5}$$

Consider now the application of this to the vector $\theta_i = (\alpha_{j,R_i})_i$ which has $d = 2^j$, with $\gamma = \gamma_j(\mathcal{F}_p^{\delta_1,\delta_2}(C))$. Then $\#\{i : |\theta_i| \ge \epsilon\} \le \min(2^j, (\gamma_j/\epsilon)^p))$. The first term 2^j is sharp for $0 \le j \le j_0$, where $j_0 = j_0(\epsilon, C; \delta_1, \delta_2)$ is the real root of $2^j = (C2^{-j(\rho+1/2-1/p)}/\epsilon)^p$. By a calculation, $j_0 = \log_2(C/\epsilon)/(\rho+1/2)$.

For notational convenience, stratify the set $\mathcal{S}$ as

$$\mathcal{S} = \mathcal{S}_0 \cup \mathcal{S}_1 \cup \cdots \mathcal{S}_j \cup \cdots,$$

where

$$\mathcal{S}_j = \{(j',R) \in \mathcal{S}, j' = j\}$$

and

$$\#\mathcal{S}_j = 2^j \quad 0 \le j \le j_0. \tag{10.6}$$

Also we have

$$\#\mathcal{S}_j \le 2^{j_0}\cdot 2^{-\beta(j-j_0)} \quad \beta = \beta(\delta_1,\delta_2,p) \qquad j \ge j_0 \tag{10.7}$$

Now consider the cover $\mathcal{S}^{**}$ defined by:

$$\begin{array}{lll} \mathcal{S}^{**} = \{ & (j,R) & 0 \le j \le j_0 \\ & (j,R) & j > j_0 \text{ and } (j,R) \text{ has a descendant in } \mathcal{S}\}. \end{array}$$

By construction, $\mathcal{S}^{**}$ contains the hereditary cover (it contains terms at $j < j_0$ which the hereditary cover might not), and so bounds on the size of $\mathcal{S}^{**}$ apply to $\mathcal{S}^*$ also. Now

$$\#\mathcal{S}^{**} \le 2^{j_0+1} + \sum_{j>j_0} A(j,R)\cdot \#\mathcal{S}_j \tag{10.8}$$

where $A(j,R)$ is the number of ancestors (j',R') of a term (j,R) at level $j_0 < j' \le j \le J$. As $A(j,R) \le (j-j_0)$,

$$\begin{aligned} \#\mathcal{S}^{**} &\le 2^{j_0+1} + \sum_{j>j_0}(j-j_0)2^{j_0}2^{-\beta(j-j_0)} \\ &= 2^{j_0}\cdot\left(2 + \sum_{j>j_0}(j-j_0)2^{-\beta(j-j_0)}\right). \end{aligned}$$

We conclude that

$$N^*(\delta_1, \delta_2, p) \leq 2^{j_0} \cdot B_1,$$

for some constant $B_1(\delta_1, \delta_2, p)$. On the other hand, by constructing a hypercube at level $\lfloor j_0 \rfloor$ using the approach of Lemmas 8.12 and 9.3, we obtain, for a constant $B_2(\delta_1, \delta_2, p)$

$$N(\delta_1, \delta_2, p) \geq B_2 \cdot 2^{j_0}.$$

Hence we may take $A = B_1/B_2$. $\square$

11 Discussion

We collect here some final remarks.

11.1 Clarifications

We would like to clearly point out that the way that the term "CART" is generally construed – as "greedy growing" of an exhaustive partition followed by "optimal pruning" in the implicit basis – is not what we have studied in this paper. Also, the data structure we have assumed – regular equispaced data on a two-dimensional rectangular lattice – is unlike the irregularly scattered data often assumed in CART studies. It would be interesting to know what properties can be established for the typical "greedy growing" non-dyadic CART algorithm, in the irregularly scattered data case.

To minimize misunderstanding, let us be clear about the intersection between CART and BOB. CART is a general methodology used for classification and discrimination or for regression. It can be used on regular or irregularly spaced data and it can construct optimal or greedy partitions within the general framework. Best-Ortho-Basis is a general methodology for adaptation of orthogonal bases to specific problems in applied mathematics. It can be used in constructing adaptive time frequency bases, and also (we have seen in this paper) in constructing adaptive bases for functions on cartesian product domains. We have shown that the methods have something in common, but, strictly speaking, only intersect under a very specific choice of problems area and entropy. Further discussion about patent lawsuits is unwarranted and pointless.

11.2 Extensions

Somewhat more general results are implicit in the results established here.

First, one can consider classes $\mathcal{F}^{\delta_1,\delta_2}_{p_1,p_2}(C_1, C_2)$ of functions obeying

$$\|D^i_h f\|_{p_i} \leq C_i h^{\delta_i}, \quad i = 1, 2.$$

The classes we have considered here in this paper are the special cases $C_1 = C_2 = C$ and $p_1 = p_2 = p$. Parallel results hold for these more general classes, and by essentially the same arguments, with a bit more book-keeping. We avoided the study of these more general classes only to simplify exposition.

Second, the log-terms we have established in Theorems 1.1 and 9.1 can be replaced by somewhat smaller log-terms. More specifically, in cases where the minimax risk scales like N^{-r} the method of proof given here actually shows that the worst-case risk of dyadic CART is within a factor $O(\log(n)^r)$ of minimax. As $0 < r < 1$, this is an improvement in the size of the log term.

11.3 Important Related Work

We also mention some related work that may be of interest to the reader.

Complexity Bounds and Oracle Inequalities. Of course there is a heavy reliance of this paper on [11, 12]. But let us also clearly point out that the general idea of oracle inequalities is clearly present in Foster and George (1995), who used a slightly different oracle less suited for our purposes here. Our underlying proof technique – the Complexity Bound underlying the proofs of Theorems 7.1 and 7.2 – is very closely related to the Minimum Complexity formalism of Barron and Cover (1991), and subsequent work by Birgé and Massart (1994).

Density Estimation. This paper grew out of a discussion with Joachim Engel, who wondered how to generalize the results of [14] to higher dimensions. Engel (personal communication) has reported progress on obtaining results on the behavior of a procedure like Dyadic CART in the setting of density estimation.

Mixed Smoothness Spaces. Neumann and von Sachs [16] have also recently studied anisotropic smoothness classes, and have shown that wavelet thresholding in a tensor wavelet basis is nearly-minimax for higher-order mixed smoothness classes. This shows that non-adaptive basis methods could also be used for obtaining nearly minimax results; the full adaptivity of CART is not really necessary for minimaxity alone.

Time-Frequency Analysis. Important related ideas are contained in two recent manuscripts associated with Coifman's group at Yale. FIrst, the manuscript of Thiele and Villemoes, which independently uses fast dyadic recursive partitioning of the kind discussed here, only in a setting where the two dimensions are time and frequency. Second, the manuscript of Bennett, which independently uses fast dyadic recursive partitioning of the kind discussed here, only in a setting where the basis functions are anisotropic Walsh functions rather than anisotropic Haar functions.

12 Appendix A: Fast Algorithms

12.1 Fast Algorithm for Dyadic CART

For a given rectangle R, we let $A(R)$ denote $Ave\{y|R\}$ and $V(R)$ denote $\mathrm{Var}\{y|R\}$. We note that if $R = R_1 \cup R_2$, then $A(R)$ and $V(R)$ are simple combinations of $A(R_i)$ and $V(R_i)$ for $i = 1, 2$. We call such combinations the *updating formulas.* The objective functional can be written

$$CPRSS(\mathcal{P}, \lambda) = \sum_{R \epsilon \mathcal{P}} \{V(R) + \lambda |R|\}. \tag{12.1}$$

Congruency Classes. Dyadic rectangles come in side lengths $2^{j1} \times 2^{j2}$ for $0 \le j_1, j_2 \le j$. We let $\mathcal{R}(j_1, j_2)$ denote the collection of all rectangles with side lengths $2^{j_1} \times 2^{j_2}$. There are $n/2^{j_1} \times n/2^{j_2}$ such rectangles.

ALGORITHM: DYADIC CART

Description. Finds the recursive dyadic partition of the $n \times n$ grid which optimizes the complexity-penalized residual sum of squares.

This algorithm, when it terminates, has placed, in the arrays *CART* and *Decor*, sufficient information to reconstruct the best partition. Indeed, starting at $R_0 = [0, n)^2$ and noting that $Decor(R_0)$ contains the indicator of the splitting direction $s(R_0)$ in Slot 2, we can follow a path to $(R^{s(R_0),0}, R^{s(R_0),1})$, and from there to their children, etc. In doing so we are traversing a decorated binary tree with decorations ["split", 1] or ["split", 2] at each nonterminal node; at the terminal nodes are the decorations $[\text{“}term'', Ave(R)]$. This describes a recursive dyadic partition $\hat{\mathcal{P}}$ and a piecewise constant reconstruction. The optimal CART value is

$$CART(R_0) \tag{12.2}$$

Inputs.

y_{i_1,i_2} $\quad 0 \le i_i, i_2 < n$. Data to be analyzed.

Results.

$(CART(R))_{R \in \mathcal{P}}$ Value of subproblem associated with R
$(Decor(R))_{R \in \mathcal{P}}$ Decoration associated with R

Internal Data Structures

$(A(R))_{R \in \mathcal{P}}$ Average of values in R
$(V(R))_{R \in \mathcal{P}}$ Sum of values in R

Initialization.

For each 1-by-1 rectangle $R_{\mathbf{i}}$
 Set $A(R_{\mathbf{i}}) = y_{\mathbf{i}}$ and $V(R_{\mathbf{i}}) = 0$

Main Loop.

For $h = 1, \cdots, 2\log_2(n)$
    For each pair (j_1, j_2) satisfying $j_1 + j_2 = h, \quad 0 \le j_1, j_2 \le h$.
        For all R in $\mathcal{R}(j_1, j_2)$
            Compute $A(R)$ from $A(R^{1,b})$, $b = 0, 1$
            Compute $V(R)$ from $A(R^{1,b}), V(R^{1,b})$, $b = 0, 1$
            Compute
                $CART(R) = \min(V(R),$
                    $CART(R^{1,0}) + CART(R^{1,1}), CART(R^{2,0}) + CART(R^{2,1}))$
            if $CART(R) \equiv V(R)$
                $Decor(R) = [\text{“term”}, A(R)]$
            Elseif $CART(R) \equiv CART(R^{1,0}) + CART(R^{1,1})$
                $Decor(R) = [\text{“split”}, 1]$
            Elseif $CART(R) \equiv CART(R^{2,0}) + CART(R^{2,1})$
                $Decor(R) = [\text{“split”}, 2]$
            End
        End
    End
End

12.2 Fast Transform into Anisotropic Haar Basis

Description. Finds the orthogonal decomposition of a vector y into linear combinations of basis elements from specified anisotropic Haar basis.

Inputs.

y_{i_1,i_2} $\quad 0 \le i_i, i_2 < n$. Data to be analyzed.
$\mathcal{R}(h)$ list of $R \in NT(\mathcal{P}^*)$ having area $n/2^h$

$s(R)$ splitting direction of $R \in NT(\mathcal{P}^*)$

Results.
$(\alpha_R)_{R \in \mathcal{P}}$ Coefficients in anisotropic Haar basis

Internal Data Structures.
$(S(R))_{R \in \mathcal{P}}$ Average of values in R $\quad$ $(V(R))_{R \in \mathcal{P}}$ Sum of values in R

Initialization.
For each 1-by-1 rectangle $R_{\mathbf{i}} = \{[i_1, i_1 + 1) \times [i_2, i_2 + 1)\}$ set $S(R_{\mathbf{i}}) = y_{\mathbf{i}}$.

For $h = 1, 2, \cdots, 2\log_2(n)$
  For each $R \in \mathcal{R}(h)$
    $S(R) = S(R^{s(R),0}) + S(R^{s(R),1})$
    $\alpha_R = (S(R^{s(R),1}) - S(R^{s(R),0})) \cdot 2^{-h/2}$
  End
End
$\beta_{[0,n)^2} = S([0, n)^2)/n$.

The algorithm takes $3N$ additions and multiplications.

12.3 Fast Algorithm for Best Anisotropic Haar Basis

Description: Finds the best orthogonal basis for the vector space of $n \times n$ arrays among all bases which arise from complete recursive dyadic partitions.

This algorithm, when it terminates, has placed, in the arrays *BOB* and *Decor*, sufficient information to reconstruct the best basis. Indeed, starting at $R_0 = [0, n)^2$ and noting that $Decor(R_0)$ contains $s(R_0)$ in Slot 2, we can follow a path to $(R^{s(R_0),0}, R^{s(R_0),1})$, and from there to their children, etc. In doing so we are building a decorated binary tree with decorations ["split", 1] or ["split", 2] at each nonterminal node; this describes a complete recursive dyadic partition $\hat{\mathcal{P}}^*$ and hence a basis $\hat{\mathcal{B}}$. Optimality of this basis follows from familiar arguments in dynamic programming. The optimal entropy is

$$\mathcal{E}(\hat{\mathcal{B}}) = e(S(R_0)|R_0|^{-1/2}) + BOB(R_0) \tag{12.3}$$

Results.
$(BOB(R))_{R \in \mathcal{P}}$ Entropy associated with R
$(Decor(R))_{R \in \mathcal{P}}$ Decoration associated with R

Internal Data Structures.
$(S(R))_{R \in \mathcal{P}}$ Sum associated with R
$(D(R, 1))_{R \in \mathcal{P}}$ Horiz Difference associated with R
$(D(R, 2))_{R \in \mathcal{P}}$ Vert Difference associated with R

Initialization.
For each $\mathbf{i} \in [0, n)^2$, set $S(R_{\mathbf{i}}) = f(\mathbf{i})$, and $BOB(R_{\mathbf{i}}) = 0$

Main Loop.

For $h = 1, 2, \cdots, 2\log_2(n)$

For each pair (j_1, j_2) with $j_i \geq 0$ and $j_1 + j_2 = h$

For each $R \in \mathcal{R}(j_1, j_2)$

$$
\begin{aligned}
S(R) &= S(R^{1,0}) + S(R^{1,1}) \\
D(R,1) &= (S(R^{1,1}) - S(R^{1,0})) \cdot 2^{-\frac{h}{2}} \\
D(R,2) &= (S(R^{2,1}) - S(R^{2,0})) \cdot 2^{-\frac{h}{2}} \\
E(R,1) &= BOB(R^{1,0}) + BOB(R^{1,1}) + e(D(R,1)) \\
E(R,2) &= BOB(R^{2,0}) + BOB(R^{2,1}) + e(D(R,2))
\end{aligned}
$$

If $E(R,1) < E(R,2)$

$Decor(R) = [\text{“split”}, 1]$

$BOB(R) = E(R,1)$

Else

$Decor(R) = [\text{“split”}, 2]$

$BOB(R) = E(R,2)$

End

End

End

End

(Note: We have written the algorithm as if every R has both vertical and horizontal splits. Actually only those R with both $j_1, j_2 \geq 1$ do so; an extra "if" branch needs to be inserted to cover the other cases).

12.4 Fast Algorithm for Best Hereditary Reconstruction

ALGORITHM: Best Heredity

Description. Finds the subtree of a given tree which optimizes the complexity-penalized residual sum of squares.

This algorithm, when it terminates, has placed, in the arrays *Best* and *Decor*, sufficient information to reconstruct the best subtree. Indeed, starting at $R_0 = [0, n)^2$ and noting that $Decor(R_0)$ contains the indicator of the splitting direction $s(R_0)$ in Slot 2, we can follow a path to $(R^{s(R_0),0}, R^{s(R_0),1})$, and from there to their children, etc. In doing so we are traversing a decorated binary tree with decorations $[\text{“split”}, 1]$ or $[\text{“split”}, 2]$ at each non-terminal node; at the terminal nodes are the decorations $[\text{“term”}, Ave(R)]$. This describes a recursive dyadic partition $\hat{\mathcal{P}}$ and a piecewise constant reconstruction. The optimal value of the CPRSS is

$$Best(R_0). \tag{12.4}$$

Inputs.

y_{i_1,i_2} $\quad 0 \leq i_i, i_2 < n$. Data to be analyzed.

$\mathcal{P}^*$ $\quad$ Complete Recursive Dyadic Partition.

Results.

$(Best(R))_{R \in \mathcal{P}^*}$ Value of subproblem associated with R

$(Decor(R))_{R \in \mathcal{P}^*}$ Decoration associated with optimal heredity

Internal Data Structures
$(A(R))_{R\in\mathcal{P}^*}$ Average of values in R
$(V(R))_{R\in\mathcal{P}^*}$ Variance of values in R

Initialization.
For each terminal rectangle R
Set $A(R) = y_{\mathbf{i}}$ and $V(R_{\mathbf{i}}) = 0$

Main Loop.
For $j = 2\log_2(n) - 1, \cdots, 1$
    For each rectangle $R \in \mathcal{R}(j)$
        Compute $A(R)$ from $A(R^{s(R),b})$.
        Compute $V(R)$ from $A(R^{s(R),b}), V(R^{s(R),b})$.
        Compute
            $Best(R) = \min\Big(V(R), Best(R^{s(R),0}) + Best(R^{s(R),1}) + \lambda\Big)$
        if $Best(R) \equiv V(R)$
            $Decor(R) = [\text{“term”}, A(R)]$
        Elseif $Best(R) \equiv Best(R^{s(R),0}) + Best(R^{s(R),1}) + \lambda$
            $Decor(R) = [\text{“split”}, s(R)]$
        End
    End
End

13 Appendix B: Proofs

13.1 Proof of Theorems 7.1 and 7.2

We prove a more general fact, concerning estimation in overcomplete dictionaries. The proof we give is a light modification of a proof in [12].

13.1.1 Constrained Minimum Complexity Estimates

Suppose we have an N by 1 vector $\mathbf{y}$ and a dictionary of N-by-1 vectors φ_γ. We wish to approximate $\mathbf{y}$ as a superposition of dictionary elements $\mathbf{y} \approx \sum_{i=1}^m \beta_\gamma \varphi_\gamma$.

We construct a matrix Φ which is N by p, where p is the total number of dictionary elements. Let each column of the Φ matrix represent one dictionary element. Note that in the case of most interest to us, $p \gg N$, as Φ “contains more than just a single basis”. For example, in the setting of this paper, $\mathcal{D}$ is the dictionary of all anisotropic Haar functions, which has approximately $p = 4N$ elements.

For approximating the vector y, we consider the vector $\tilde{\beta} \in R^p$, the vector $\tilde{f} = \Phi\tilde{\beta}$ denotes a corresponding linear combination of dictionary elements. This places the approximation $\tilde{f}$ in correspondence with the coefficient vector $\tilde{\beta}$. Owing to the possible overcompleteness of Φ, this correspondence is in general one-to-many.

Define now the *empirical complexity* functional

$$K(\tilde{f}, y) = \|\tilde{f} - y\|_2^2 + \lambda^2 \cdot \sigma^2 \cdot N(\tilde{f}),$$

where

$$N(\tilde{f}) = \min_{\tilde{f}=\Phi\tilde{\beta}} \#\{j : \tilde{\beta}_j \neq 0\}$$

is the complexity of constructing $\tilde{f}$ from the dictionary Φ. Also, define the *theoretical complexity* functional

$$K(\tilde{f}, f) = \|\tilde{f} - f\|_2^2 + \lambda^2 \cdot \sigma^2 \cdot N(\tilde{f}).$$

Let $\mathcal{C}$ be a collection of "allowable" coefficient vectors $\beta \in R^p$. We will be interested in approximations to y obeying these constraints and having small complexity. In a general setting, one can think of many interesting constraints to impose on allowable coefficients; for example, that coefficients should be positive, that coefficients should generate a monotone function, that nonzero coefficients are attached to pairwise orthogonal elements.

Define the $\mathcal{C}$-constrained minimum empirical complexity estimate

$$\hat{f}^* = \operatorname*{argmin}_{\{\tilde{f}=\Phi\tilde{\beta}:\tilde{\beta}\in\mathcal{C}\}} K(\tilde{f}, y).$$

In a moment we will prove the

Complexity Bound. Suppose $y = f + z$, where z is i.i.d. $N(0,1)$. Fix $\mathcal{C} \subset R^p$ and fix $\zeta > 8$, and consider the $\mathcal{C}$-constrained minimum-complexity model selection with $\lambda = \zeta \cdot (1 + \sqrt{2 \log p})$.

$$E\ K(\hat{f}^*, f) \leq A(\zeta) \cdot \left(\lambda^2 \sigma^2 + \min_{\{\tilde{f}=\Phi\tilde{\beta}:\tilde{\beta}\in\mathcal{C}\}} K(\tilde{f}, f) \right). \tag{13.1}$$

This shows that the empirical minimum complexity estimate is not far off from minimizing the theoretical complexity.

13.1.2 Relation to CART and BOB

We now explain why the complexity bound implies Theorems 7.1 and 7.2.

We begin with the observation that the empirical complexity $K(f, y)$ is just what we earlier called a complexity-penalized sum of squares.

Assume now that the dictionary $\mathcal{D}$ is the collection of all anisotropic Haar functions. Two constraint sets are particularly interesting.

First, let $\mathcal{C}_{BOB}$ be the collection of all coefficient vectors β which arise from combinations of atoms that all belong together in some ortho basis built from the Anisotropic Haar dictionary. Remember, the dictionary has $p \gg N$ atoms. So at most N elements of β can be nonzero at once under this constraint. Also, we have seen in section 3 that each basis in the anisotropic Haar system corresponds to a certain decorated tree so this constraint says that collections of coefficients which are allowed to be nonzero simultaneously correspond to certain collections of indices. This constaint can be made quite explicit and algorithmic, although we do not go into details here.

If we optimize the empirical complexity $K(\tilde{f}, y)$ over all $\tilde{f}$ arising from a $\beta \in \mathcal{C}_{BOB}$ we get exactly the estimator (4.1)-(4.3). We encourage the reader to check this fact.

Second, there is the CART constraint. Let $\mathcal{C}_{CART}$ be the collection of all vectors β for which the nonzero coefficients in the corresponding β only refer to atoms which can appear together in an orthogonal basis, and fow which the nonzero coefficients only occur in an hereditary pattern in that basis. We remark that $\mathcal{C}_{CART} \subset \mathcal{C}_{BOB}$.

If we optimize the empirical complexity $K(\tilde{f}, y)$ over all $\tilde{f}$ arising from a $\beta \in \mathcal{C}_{CART}$ we get exactly the estimator (2.4)-(2.5). We again encourage the reader to check this fact.

We now make two simple observations about the minimum complexity formalism, valid for any $\mathcal{C}$, which the reader should verify:

K1. The theoretical complexity of $\hat{f}^*$ upperbounds the predictive loss:

$$K(\hat{f}^*, f) \geq \|\hat{f}^* - f\|_2^2.$$

K2. The minimum theoretical complexity is within a logarithmic factor of the ideal risk:

$$\begin{aligned} \min_{\tilde{f}\in\mathcal{C}} K(\tilde{f}, f) &= \min_{\tilde{f}\in\mathcal{C}} \|\tilde{f} - f\|_2^2 + \lambda^2 \cdot \sigma^2 \cdot N(\tilde{f}) \\ &\leq \lambda^2 \cdot \min_{\tilde{f}\in\mathcal{C}} (\|\tilde{f} - f\|_2^2 + \sigma^2 \cdot N(\tilde{f})) \\ &= \lambda^2 \cdot \min_{\tilde{\beta}\in\mathcal{C}} (\|\Phi\tilde{\beta} - \Phi\beta\|_2^2 + \sigma^2 \cdot \#\{j : \tilde{\beta}_j \neq 0\}) \\ &= \lambda^2 \cdot \mathcal{R}(Ideal\mathcal{C}, f). \end{aligned}$$

These observations, translated into the cases $\mathcal{C}_{BOB}$ and $\mathcal{C}_{CART}$, give Theorems 7.1 and 7.2 respectively.

13.1.3 Proof of the Complexity Bound

In what follows we assume the noise level $\sigma^2 = 1$. We follow Donoho and Johnstone (1995) line-by-line, who analyzed the unconstrained case $\mathcal{C} = R^p$. Exactly the same analysis applies in the constrained case.

We first let f^0 denote a model of minimum theoretical complexity:

$$K(f^0, f) = \min_{\tilde{f}\in\mathcal{C}} K(\tilde{f}, f).$$

As $\hat{f}^*$ has minimum empirical complexity,

$$K(\hat{f}^*, y) \leq K(f^0, y).$$

As $\|\hat{f}^* - y\|_2^2 = \|\hat{f}^* - f - z\|_2^2$ we can relate empirical and theoretical complexities by

$$K(\hat{f}^*, y) = K(\hat{f}^*, f) + 2\langle z, f - \hat{f}^*\rangle + \|z\|_2^2,$$

and so combining the last two displays,

$$K(\hat{f}^*, f) \leq K(f^0, f) + 2\langle z, \hat{f}^* - f^0\rangle.$$

Now define the random variable

$$\mathcal{W}(k) = \sup\{\langle z, m^2 - m^1\rangle : \|m^j - f\|_2^2 \leq k, \lambda^2 N(m^j) \leq k\}.$$

Then

$$K(\hat{f}^*, f) \leq K(f^0, f) + 2\mathcal{W}(K(\hat{f}^*, f)).$$

This display shows the key idea. It turns out that $\mathcal{W}(k) \ll k$ for all large k, and so this display forces $K(\hat{f}^*, f)$ to be not much larger than $K(f^0, f)$.

Denote the minimum theoretical complexity by $K^0 = K(f^0, f)$. Define $k_j = 2^j \cdot (1 - 8/\zeta)^{-1} \cdot \max(K^0, \lambda^2)$ for $j \geq 0$. Define the event

$$B_j = \{\mathcal{W}(k) \leq 4k/\zeta : \quad \text{for all } k \geq k_j\}.$$

On the event B_j, the inequality

$$k \leq K^0 + 2\mathcal{W}(k)$$

has no solutions for $k \geq k_j$. Hence, on event B_j,

$$K(\hat{f}^*, f) \leq k_j.$$

It follows that

$$\begin{aligned} EK(\hat{f}^*, f) &\leq \sum_{j=0}^{\infty} k_{j+1} Prob\{K(\hat{f}^*, f) \in [k_j, k_{j+1})\} \\ &\leq \sum_{j=0}^{\infty} k_{j+1} Prob\{K(\hat{f}^*, f) \geq k_j\} \\ &\leq \sum_{j=0}^{\infty} k_{j+1} Prob\{B_j^c\}. \end{aligned}$$

By Lemma 13.1 we get

$$\begin{aligned} EK(\hat{f}^*, f) &\leq k_0 \cdot \sum_{j \geq 0} 2^{j+1}/(2^j)! \\ &\leq \max(K^0, \lambda^2) \cdot (1 - 8/\zeta)^{-1} \cdot 6. \end{aligned}$$

Hence, the complexity bound (13.1) holds, with $A(\zeta) = (1 - 8/\zeta)^{-1} \cdot 6$.

Lemma 13.1 *(Donoho and Johnstone, 1994)*

$$Prob\{B_j^c\} \leq 1/(2^j)!.$$

The proof depends on tail bounds for Chi-Squared variables, which, ultimately, depend on concentration-of-measure estimates (e.g. Borell-Tsirel'son inequality).

13.2 Proof of Lemma 8.4

The proof of each display is similar, so we just discuss the first. Fix a rectangle $R = I_1 \times I_2$ with $|I_1| = 2^{-j_1}$. Let $R^{1,0}$ and $R^{1,1}$ denote the left and right halves.

$$\langle f, \phi_R^1 \rangle = 2^{j/2} \left\{ \int_{R^{1,1}} f dx dy - \int_{R^{1,0}} f dx dy \right\}.$$

Hence for the very special increment $h = 2^{-j_1}$

$$\begin{aligned} \int_{R^{1,0}} (D_h f)(x, y) dx dy &= \int_{R^{1,0}} (f(x+h, y) - f(x, y)) dx dy \\ &= \int_{R^{1,1}} f dx dy - \int_{R^{1,0}} f dx dy = 2^{-j/2} \langle f, \phi_R^1 \rangle. \end{aligned}$$

For any sum $\sum_R$ over rectangles R with disjoint interiors,

$$\sum_R |\langle f, \phi_R^1\rangle|^p = 2^{jp/2} \sum_R |\int_{R^{1,0}} D_h^1 f|^p.$$

Now if $1/p + 1/p' = 1$,

$$|\int_{R^{1,0}} D_h^1 f| \leq \|D_h^1 f\|_{L^p(R^{1,0})} \|1\|_{L^{p'}(R^{1,0})},$$

so

$$\sum_R |\langle f, \phi_R^1\rangle|^p \leq 2^{j(p/2+(1-p)} \cdot \sum_R \|D_h^1 f\|_{L^p(R^{1,0})}^p.$$

Now if $\sum_R$ is interpreted to mean the sum over a partition of $[0,1]^2$ by congruent rectangles, then

$$\sum_R \|D_h^1 f\|_{L^p(R)}^p = \|D_h^1 f\|_{L^p(Q_h^1)}^p,$$

and so from $\|D_h^1 f\|_{L^p(R^{1,0})} \leq \|D_h^1 f\|_{L^p(R)}$ we conclude that

$$\begin{aligned} (\sum_{R \in \mathcal{R}(j)} |\langle f, \phi_R^1\rangle|^p)^{1/p} &\leq 2^{j(1/p-1/2)} \|D_h^1 f\|_{L^p(Q_h^1)} \\ &\leq 2^{j(1/p-1/2)} \cdot h^{\delta_1} \cdot C = 2^{-j_1\delta_1} \cdot 2^{j(1/p-1/2)} \cdot C. \end{aligned}$$

13.3 Proof of Lemma 8.6

We assume for the proof below that δ_1 and δ_2 are mutually irrational. Very slight modifications allow to handle the exceptional cases.

Think of the quarterplane consisting of (x, y) with $x, y \geq 0$ as a collection of square "unit cells", with vertices on the integer lattice. Think of the set where $x\delta_1 = y\delta_2$ as a ray S in this quarterplane, originating at $(0,0)$.

Let $p_j = (j_1(j), j_2(j))$ denote the sequence of pairs of values $j_1\delta_1 = y\delta_2$ where $j_1 + j_2 = j$. Let $p_j^* = (j_1^*(j), j_2^*(j))$ denote the sequence of pairs of values obtained from the optimal sequential partitioning of definition 8.5.

Our Claim, to be established below: *p_j and p_j^* always belong to the same unit cell.*

It follows from this Claim that $j_1^*(j) > j_1(j) - 1$ and $j_2^*(j) > j_2(j) - 1$; as a result

$$\max(2^{-j_1^*(j)\delta_1}, 2^{-j_2^*(j)\delta_1}) \leq 2 \cdot \max(2^{-j_1(j)\delta_1}, 2^{-j_2(j)\delta_2}),$$

and the lemma follows.

The Claim is proved by induction. Indeed, at $j = 0$, $p_j = p_j^* = 0$. So the claim is true at $j = 0$.

For the inductive step, suppose the Claim is true for steps $0 \leq j \leq J$; we prove it for $J + 1$. Let C_j denote "the" unit cell containing p_j, where, if several cells qualify, we select a cell having p_j on the skew diagonal.

Under this convention, at each step j, p_j lies on the skew diagonal through this cell which joins its upper left corner to its lower right corner. Supposing the Claim is true at step j, p_j^* is either at the upper left corner or at the lower right corner of the cell. Note also that C_{j+1} is either above C_j, or to the right of C_j.

With this set-up, the inductive step requires two things: (i) that if p_J^* is at the lower right corner of C_J, and C_{J+1} is above C_J, then, p_{J+1}^* is above p_J^* i.e. $b_2(J+1) = 1$; (ii) that if p_J^* is at the upper left corner of C_J, and if C_{J+1} is the cell to the right of C_J, then p_{J+1}^* is to the right of p_J^* i.e. $b_1(J+1) = 1$.

Now note that the trajectory of p_J^* is being determined by greedy minimization of the function $f(x,y) = \max(2^{-\delta_1 x}, 2^{-\delta_2 y})$ by paths through integer lattice points. Below the ray S, $\frac{\partial}{\partial x} f(x,y) = 0$. We conclude that unit moves in the x-direction are useless when one is below S. On the other hand, below S, $\frac{\partial}{\partial y} f(x,y) < 0$. So a unit move in the y-direction, *if it is available* is useful. Above the ray S, the situation is reversed: $\frac{\partial}{\partial y} f(x,y) = 0$. We conclude that any move in the y-direction is useless when one is above S. But a unit move in the x-direction, *if it is available* is useful.

Suppose one is in case (i) of the above paragraph. Then one knows that the upper right vertex of C_J is below or on the ray S. It follows that a full unit move in the y direction is available and useful. The greedy algorithm will certainly take it, and case (i) is established.

Suppose one is in case (ii) of the above paragraph. Then one knows that the upper right vertex of C_J is above or on the ray S. It follows that a full unit move in the x direction is both available and useful. The greedy algorithm will certainly take it, and case (ii) is established.

13.4 Proof of Lemma 8.10

Define

$$N(\epsilon) = \sup\{\#\{i : |\theta_i(f, BAB(\delta_1,\delta_2))| \geq \epsilon\} : f \in \mathcal{F}_p^{\delta_1,\delta_2}(C)\}.$$

The property in question amounts to the assertion that

$$N(\epsilon)^{\rho+1/2}\epsilon \leq K \cdot C, \qquad \forall \epsilon > 0. \tag{13.2}$$

By Corollary 8.7, there are constants $\gamma_j = \gamma_j(C)$ so that for $f \in \mathcal{F}_p^{\delta_1,\delta_2}(C)$, the coefficients in $BAB(\delta_1,\delta_2)$ obey

$$(\sum_{\mathcal{R}(j)} |\alpha_R|^p)^{1/p} \leq \gamma_j.$$

Now define

$$n(\epsilon, d, \gamma) = \sup\{\#\{i : |\theta_i| \geq \epsilon\} : \theta \in R^d, \|\theta\|_{\ell^p} \leq \gamma\}.$$

Then

$$N(\epsilon) \leq 1 + \sum_{j \geq 0} n(\epsilon, 2^j, \gamma_j),$$

where the γ_j are as above. Easy calculations (see (10.4) and (10.5)) yield $n(\epsilon, d, \gamma) = \min(d, (\gamma/\epsilon)^p)$; from $\gamma_j = C2^{-j(\rho+1/2-1/p)}$ we get (13.2).

13.5 Proof of Lemma 8.11

The proof is an application of the following fact, called the "incompressibility of Hypercubes" in [8]. Suppose that $\mathcal{H}$ is an orthogonal hypercube symmetric about zero; then it can be written $\sum_j \alpha_j g_j$ where the g_j are orthogonal and the α vary throughout the cube $|\alpha_j| \leq \epsilon$. We call any basis starting with elements $g_1, g_2, \ldots, g_m$ a natural basis for $\mathcal{H}$. In that basis, $\mathcal{H}$ is rotated so that the axes cut orthogonally through its faces.

Let $\mathcal{B}$ be a natural basis for such a $\mathcal{H}$ and let $\Theta = \Theta(\mathcal{H},\mathcal{B})$ be the body of coefficients of $\mathcal{H}$ in that basis. Let U be any orthogonal matrix. Then for absolute constants $c(p)$, and $0 < p \leq 2$

$$\sup_{\theta\in\Theta} \|U\theta\|_{l^p} \geq c(p) \sup_{\theta\in\Theta} \|\theta\|_{l^p}.$$

In [8], this is shown to be a consequence of Khintchine's inequality

To use this, we argue by contradiction. Suppose that the hypotheses of the Lemma hold, and yet for a certain basis $\mathcal{B}^*$, $\tau^*(\Theta(\mathcal{F},\mathcal{B}^*)) = \tau - \epsilon$ where $\epsilon > 0$. Then for $0 < \delta < \epsilon$, we have the weak-type inclusion $\Theta(\mathcal{F},\mathcal{B}^*) \subset wl^{\tau-\delta}$. Equally, we have the stronger inclusion $\Theta(\mathcal{F},\mathcal{B}^*) \subset l^{\tau-\delta}$.

Let $\mathcal{H}_j$ be the j-th hypercube in the sequence posited by the Theorem, and let $\mathcal{B}_j$ be a natural basis for $\mathcal{H}_j$. There is an orthogonal matrix U_j so that $\theta(f,\mathcal{B}^*) = U_j\theta(f,\mathcal{B}_j)$.

$$\begin{aligned}
\sup_{h\in\mathcal{H}_j} \|\theta(h,\mathcal{B}^*)\|_{l^{\tau-\delta}} &= \sup_{h\in\mathcal{H}_j} \|U_j\theta(h,\mathcal{B}_j)\|_{l^{\tau-\delta}} \\
&\geq c(\tau-\delta) \sup_{h\in\mathcal{H}_j} \|\theta(h,\mathcal{B}_j)\|_{l^{\tau-\delta}}
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\sup_{h\in\mathcal{H}_j} \|\theta(h,\mathcal{B}_j)\|_{l^{\tau-\delta}} &= m_j^{1/\tau-\delta}\epsilon_j \\
&= (m_j^{1/\tau}\epsilon_j) m_j^{1/(\tau-\delta)-1/\tau} \\
&\geq c_0 m_j^{1/(\tau-\delta)-1/\tau} \to \infty.
\end{aligned}$$

Hence $\Theta(\mathcal{F},\mathcal{B}^*) \not\subset l^{\tau-\delta}$ for any $\delta > 0$. This contradiction proves the Lemma.

13.6 Proof of Lemma 8.12

The Construction. Let g be a smooth function on R^2 supported inside the unit square $[0,1]^2$, whose support contains the half-square $[1/2,3/4]^2$. Suppose that $\|\frac{\partial}{\partial x}g\|_{L^\infty} \leq \xi$ and that $\|\frac{\partial}{\partial y}g\|_{L^\infty} \leq \xi$. Suppose also that $\|g\|_{L^2} = 1$.

Let $\mathcal{R}(j)$ be the tiling of $[0,1]$ selected at level j by $BAB(\delta_1,\delta_1)$. As this is a spatially homogeneous basis, all tiles are congruent. For an $R \in \mathcal{R}(j)$, let g_R denote the translation and dilation of g so that it "just fits" inside R: i.e. $supp(g_R) \subset R$ and $R/2 \subset supp(g_R)$ where $R/2$ denotes the rectangle with the same center homothetically shrunk by a factor of 50%.

Let $\epsilon_j = \frac{1}{6\xi}C2^{-j(\rho+1/2)}$, and set

$$\mathcal{H}_j = \sum_{R\in\mathcal{R}(j)} \alpha_R g_R, \qquad \text{where } |\alpha_R| \leq \epsilon_j.$$

Property 1. We first note that $\mathcal{H}_j$ obeys the dimension inequality assumed in the statement of the lemma, with $K = 1/6\xi$. Set $\rho = \delta_1\delta_2/(\delta_1+\delta_2)$ and $\tau = 1/(2\rho+1)$. With $m_j = 2^j$ the dimension of $\mathcal{H}_j$ and ϵ_j the sidelength, one gets

$$m_j^{1/\tau}\epsilon_j = C_0 > 0$$

with $C_0 = C/6\xi$.

Property 2. The key claim about $\mathcal{H}_j$ is the embedding $\mathcal{H}_j \subset \mathcal{F}_p^{\delta_1,\delta_2}(C)$: for any $f \in \mathcal{H}_j$,

$$\sup_{0<h<1} h^{-\delta_1} ||D_h^1 f||_{L^p(Q_h^1)} \le C \tag{13.3}$$

$$\sup_{0<h<1} h^{-\delta_2} ||D_h^2 f||_{L^p(Q_h^2)} \le C$$

We prove the first inequality only, starting with estimates for differences of g_R. Let R be of side 2^{-j_1} by 2^{-j_2}.

Let $h > 2^{-j_1}$, and let R_h denote the translation of R by "to the left by h". Then if $(x,y) \in R_h$, $D_h^1 g_R(x,y) = g_R(x+h,y)$, while if $(x,y) \in R$, $D_h^1 g_R(x,y) = -g_R(x,y)$. Note further that R_h is not generally part of the tiling $\mathcal{R}(j)$, but instead overlaps with two tiles, R_h^- and R_h^+, say. Let $b_R(x,y) = g_R(x+h,y)1_{R_h^-}(x,y)$, and $c_R(x,y) = g_R(x+h,y)1_{R_h^+}(x,y)$. Then $D_h^1 g_R = a_R + b_r + c_R$, where a_R is supported in R, and b_R and c_R are supported in $R_h^{\pm}$. We have for each R

$$\|a_R\|_\infty, \|b_R\|_\infty, \|c_R\|_\infty \le \xi 2^{j/2}.$$

Now consider the case $0 < h \le 2^{-j_1}$. Let $b_R(x,y) = g_R(x+h,y)1_{R_h}(x,y)$, and $c_R(x,y) = 0$ and set $R_h^+ = R_h^- = R_h$. Then $D_h^1 g_R = a_R + b_r + c_R$, where a_R is supported in R, and b_R and c_R are supported in $R_h^{\pm}$. We have

$$\|a_R\|_\infty, \|b_R\|_\infty \le \min(h2^{j_1}, 1)\xi 2^{j/2}.$$

Now consider increments of $f = \sum_R \alpha_R g_R$. Rearrange the terms to have common support:

$$D_h^1 f = \sum_R \alpha_R a_R + \alpha_{R_h^-} b_R + \alpha_{R_h^+} c_R.$$

Now

$$\|(\alpha_{R_h^-})\|_{\ell^p}, \|(\alpha_{R_h^+})\|_{\ell^p} \le \|(\alpha_R)\|_{\ell^p}$$

and so

$$\begin{aligned} \|D_h^1 f\|_p &\le \|\sum_R \alpha_R a_R\|_p + \|\sum \alpha_{R_h^-} b_R\|_p + \|\sum_R \alpha_{R_h^+} c_R\|_p \\ &\le \|(\alpha_R)\|_{\ell^p} \max_R \|a_R\|_\infty + \|(\alpha_{R_h^-})\|_{\ell^p} \max_R \|b_R\|_\infty + \|(\alpha_{R_h^+})\|_{\ell^p} \max_R \|c_R\|_\infty \\ &\le 3\xi \min(h2^{j_1}, 1)\|(\alpha_R)\|_{\ell^p}. \end{aligned}$$

Hence

$$\begin{aligned} \sup_{0<h<1} h^{-\delta_1} \|D_h^1 f\|_p &\le \sup_h h^{-\delta_1} 3\xi \|(\alpha_R)\|_{\ell^p} \min(h2^{j_1}, 1) 2^{j(1/2-1/p)} \\ &= 3\xi \|(\alpha_R)\|_{\ell^p} 2^{j(1/2-1/p)} \sup_h h^{-\delta_1} \min(h2^{j_1}, 1) \\ &= 3\xi \|(\alpha_R)\|_{\ell^p} 2^{j(1/2-1/p)} 2^{j_1\delta_1} \end{aligned}$$

Now from the proof of Lemma 8.6 we know that for $BAB(\delta_1, \delta_2)$,

$$2^{j_1\delta_1} \le 2 \cdot 2^{j\rho}$$

we conclude that

$$\sup_{0<h<1} h^{-\delta_1} \|D_h^1 f\|_p \le 6\xi 2^{j(\rho+1/2-1/p)} \|\alpha\|_p \le C$$

This establishes (13.3).

13.7 Proof of Lemma 9.3

Recall the proof of Lemma 8.12. Let $\epsilon > 0$ be given, and pick j so that the values $\epsilon_j, \epsilon_{j+1}$ defined in that lemma satisfy

$$\epsilon_{j+1} < \epsilon \leq \epsilon_j.$$

Construct the Hypercube $\mathcal{H}_j$ exactly as in that Lemma, only using sidelength ϵ in place of ϵ_j.

We first note that the generating elements g_R are orthogonal with respect to the sampling measure ℓ^2_N, because they are disjointly supported. We also note that because of the dyadic structure of the sampling and the congruency of the hypercubes, each g_R has the same ℓ^2_N norm as every other g_R. Call this norm $\delta(\epsilon)$. Finally, we note that

$$\delta = \epsilon \cdot (\frac{1}{M_1 M_2} \sum g(x_i)^2)^{1/2} / ||g||_{L^2[0,1]^2},$$

where the sum is over an M_1 by M_2 array of grid points, where $M_i = 2^{J - j_i(j)}$. Hence $\mathcal{H}_j$ is an orthogonal hypercube for ℓ^2_N. The asymptotics of the sidelength can be derived from the fact that g is nice, the grid is becoming finer as j increases, and so the indicated sum converges to the corresponding integral, whence

$$\delta = \epsilon \cdot (1 + o(1)).$$

The Hypercube $\mathcal{H}_j$ that results has two properties: First,

$$m_j^{1/\tau} \epsilon = C_1(\epsilon)$$

where

$$C_1(\epsilon) = C_0 \cdot (\epsilon/\epsilon_j) > C_0 \cdot (\epsilon_{j+1}/\epsilon_j) = C/6\xi 2^{-(\rho+1/2)}.$$

Hence the dimensionality of the hypercube obeys (9.4), with $K = (6 \cdot \xi \cdot 2^{(\rho+1/2)})^\tau$.

Second,

$$\mathcal{H}_j \subset \mathcal{F}_p^{\delta_1, \delta_2}(C).$$

This inclusion follows exactly as in Lemma 8.12.

References

[1] A. Barron and T. Cover (1991) Minimum Complexity Density estimation. *IEEE Trans. Info. Theory* **37**, 1034-1054.

[2] N. Bennett. Computing Best Dyadic Local Cosine and Wavelet Packet Bases. Manuscript, 1995.

[3] L. Birgé and P. Massart (1994) From Model Selection to Adaptive Estimation. To appear, Le Cam Festschrift, D. Pollard and G. Yang eds.

[4] L. Breiman, J. Friedman, R. Olshen, and C.J Stone. *Classification and Regression Trees.* Belmont, CA: Wadsworth, 1983.

[5] R. R. Coifman, Y. Meyer, S. Quake, and M. V. Wickerhauser, "Wavelet analysis and signal processing", pp. 363-380 in *Wavelets and Their Applications*, J.S. Byrnes, J.L. Byrnes, K.A. Hargreaves, and K. Berry, eds., Kluwer Academic: Boston, 1994.

[6] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best-basis selection", *IEEE Trans. Info. Theory*, **38**, pp. 713-718, 1992.

[7] D.L. Donoho, "De-Noising by Soft Thresholding", *IEEE Trans. Info. Thry.*, **41**, pp. 613-627, 1995.

[8] D.L. Donoho, *Unconditional bases are optimal bases for data compression and for statistical estimation*, Applied and Computational Harmonic Analysis, **1**, 100-115, 1993.

[9] D.L. Donoho, "Abstract Statistical Estimation and Modern Harmonic Analysis", *Proc. 1994 Int. Cong. Math.*, to appear 1995.

[10] D.L. Donoho and I.M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage", *Biometrika*, vol. 81, pp. 425-455, 1994.

[11] D.L. Donoho and I.M. Johnstone, *Ideal De-Noising in a basis chosen from a library of orthonormal bases*, Comptes Rendus Acad. Sci. Paris A, **319**, 1317-1322, 1994.

[12] D.L. Donoho and I.M. Johnstone, "Empirical Atomic Decomposition", Manuscript.

[13] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard, "Wavelet Shrinkage: Asymptopia?", *Journ. Roy. Stat. Soc.* Ser **B**, vol. 57, no. 2, pp. 301-369, 1995.

[14] J. Engel, A simple wavelet approach to nonparametric regression from recursive partitioning schemes, J. Multivariate Analysis **49** $242-254$, 1994.

[15] D. Foster and E.I. George, The risk inflation factor in multiple linear regression. *Ann. Statist.* 1995.

[16] M. Neumann and R. von Sachs, Anisotropic wavelet smoothing, with an application to nonstationary spectral analysis. Mss, 1994.

[17] V. Temlyakov, Approximation of functions with bounded mixed derivatives, Trudy Mat. Inst. Akad. Nauk. SSSR, **178**, 1-112, 1986.

[18] C.M. Thiele and L.F. Villemoes (1994) A fast algorithm for Adapted Time-Frequency Tilings. Manuscript.