The Annals of Statistics
1999, Vol. 27, No. 3, 859–897

WEDGELETS: NEARLY MINIMAX ESTIMATION OF EDGES[1]

BY DAVID L. DONOHO

Stanford University

Dedicated to Rafail Khas'minskii in his 65th year

We study a simple "horizon model" for the problem of recovering an image from noisy data; in this model the image has an edge with α-Hölder regularity. Adopting the viewpoint of computational harmonic analysis, we develop an overcomplete collection of atoms called *wedgelets*, dyadically organized indicator functions with a variety of locations, scales and orientations. The wedgelet representation provides nearly optimal representations of objects in the horizon model, as measured by minimax description length.

We show how to rapidly compute a wedgelet approximation to noisy data by finding a special *edgelet-decorated* recursive partition which minimizes a complexity-penalized sum of squares. This estimate, using sufficient subpixel resolution, achieves nearly the minimax mean-squared error in the horizon model. In fact, the method is adaptive in the sense that it achieves nearly the minimax risk for any value of the unknown degree of regularity of the horizon, $1 \le \alpha \le 2$.

Wedgelet analysis and denoising may be used successfully outside the horizon model. We study images modelled as indicators of star-shaped sets with smooth boundaries and show that complexity-penalized wedgelet partitioning achieves nearly the minimax risk in that setting also.

1. Introduction. Consider a simple mathematical model of the problem of removing noise from image data. We are interested in an object $f(x_1, x_2)$ defined on the unit square $S = \{(x_1, x_2): 0 \le x_1, x_2 \le 1\}$ and we assume we have available a means of measuring pixel-level averages about f with noise. Formally, we let $\text{Pixel}(i_1, i_2) = [i_1/n, (i_1 + 1)/n) \times [i_2/n, (i_2 + 1)/n)$, and we assume we are able to get noisy measurements of the pixel-level averages $\tilde{f}(i_1, i_2) = \text{Ave}\{f \mid \text{Pixel}(i_1, i_2)\}$. Thus we have available an n-by-n array of data

$$y_{i_1, i_2} = \tilde{f}(i_1, i_2) + z_{i_1, i_2}, \qquad 0 \le i_1, i_2 < n, \tag{1.1}$$

where the $z_{i_1, i_2} \sim_{\text{iid}} N(0, \sigma^2)$ are samples from a white Gaussian noise. We wish to recover f with small per-pixel mean-squared error $\text{MSE}(\hat{f}, f) = En^{-2} \sum_{(i_1, i_2)} (\hat{f}(i_1, i_2) - \tilde{f}(i_1, i_2))^2$. We take a minimax point of view, defining

Received September 1997; revised May 1999.

[1]Research supported in part by NSF Grant DMS-95-05151, AFOSR Grant MURI95-F49620-96-1-0028 and by other sponsors.

AMS 1991 *subject classifications*. Primary 62G07, 62C20; secondary 62G20, 41A30, 41A63.

Key words and phrases. Minimax estimation, edges, edgels, edgelets, fast algorithms, complexity penalized estimates, recursive partitioning, subpixel resolution, oracle inequalities.

a function class $\mathscr{F}$ and searching for estimators exactly or approximately attaining the minimax mean-squared error:

$$M^*(n, \mathscr{F}) = \inf_{\hat{f}} \sup_{\mathscr{F}} \mathrm{MSE}(\hat{f}, f). \tag{1.2}$$

This type of minimax estimation problem has been studied at length in many papers in the literature, of course, under the assumption that $\mathscr{F}$ is a smoothness class such as $\{f: \|f^{(m)}\|_{L^p} \le C\}$; then it is properly a problem of "smoothing."

1.1. *Edges in images.* We are interested in the problem of dealing with objects f that are discontinuous along curves: this is a way of directing our attention to the fact that, in real-world image data, the most interesting aspects of the image are the edges. The importance of edges in the vision literature goes back to the work of Marr [32] and even earlier; one could say that this issue permeates the field. In the statistical literature pioneers in bringing the "edge" issue to the fore include Khas'minskii and Lebedev [27] and Korostelev and Tsybakov [29], whose initial efforts were roughly synchronous. There is also interesting work by Geman and Geman [25] and by Müller and Song [35] oriented to a different set of goals.

Consider the following very simple "horizon" model. Suppose there is a function $H(x)$ called the horizon, defined on the interval $[0, 1]$, and that the image is of the form

$$f(x_1, x_2) = 1_{\{x_2 \ge H(x_1)\}}, \qquad 0 \le x_1, x_2 \le 1.$$

This models a "black-and-white image" with a horizon, where the image is "white" above the horizon and "black" below. We are interested in cases where the horizon is regular, and to measure this we use Hölder conditions. For $0 < \alpha \le 1$ we say that $H \in \text{HÖLDER}^\alpha(C)$ if

$$|H(x) - H(x')| \le C \cdot |x - x'|^\alpha, \qquad 0 \le x, x' \le 1.$$

For $1 < \alpha \le 2$ we say that $H \in \text{HÖLDER}^\alpha(C)$ if

$$|H'(x) - H'(x')| \le C \cdot |x - x'|^{(\alpha - 1)}, \qquad 0 \le x, x' \le 1,$$

where H' is the derivative of H. For $\alpha = 1$ ($\alpha = 2$) membership in HÖLDER^α imposes a Lipschitz condition on H (respectively, on H'); for $0 < \alpha < 1$ we are measuring a degree of fractional regularity of H, and for $1 < \alpha < 2$ a degree of fractional regularity of H'. We define a functional class $\text{HORIZ}^\alpha(C_1, C_\alpha)$,

$$\text{HORIZ}^\alpha(C_1, C_\alpha) = \{f: H \in \text{HÖLDER}^\alpha(C_\alpha) \cap \text{HÖLDER}^1(C_1)\}. \tag{1.3}$$

This model is essentially the *model of boundary fragments* introduced by Korostelev and Tsybakov [29]; we find the name *horizon model* more evocative.

We will focus in this paper on the recovery of images in model (1.1) under the performance criterion (1.2) with $\mathscr{F} = \text{HORIZ}^\alpha(C_1, C_\alpha)$.

There are of course many ways that one could choose to treat data (1.1) arising from model (1.3). Since the image is locally constant, traditional smoothing techniques (kernel methods) could be applied; these would not make use of the specific structure of the image as something where "all the action is in the edges." At the other extreme, one could develop estimators very specifically tailored to "edge finding," perhaps also exploiting the "black-or-white" aspect of the image directly.

1.2. *Computational harmonic analysis.* In this paper, we are interested in exploring the point of view of computational harmonic analysis (CHA), a rapidly developing discipline whose recent achievements include the development of wavelets, wavelet packets, cosine packets, brushlets and other novel schemes of data representation [7, 9, 30, 34, 38]. There is an emerging tradition within CHA, whereby the "way to go about things" is to find the "optimal representation" of the objects underlying a problem and then a "fast algorithm" to compute that representation.

The CHA viewpoint says that the "optimal representation" depends on the functional class, and that once one has the "optimal representation" for a functional class $\mathscr{F}$ one can easily do many different tasks. One of those tasks is to remove noise (i.e., construct minimax estimators); another task is to work with noiseless data and perform data compression. The CHA point of view would say that "optimal representation" is primary, and that statistical and information theoretic applications follow directly [12, 15].

From this very general point of view, a number of asymptotic minimaxity results in mathematical statistics are seen as special cases of a larger picture. Among those results we identify the following:

1. the fact that Fourier series estimates are nearly minimax over L^2-Sobolev classes;
2. the fact that wavelet estimates are nearly minimax over L^p-Sobolev, Hölder, Triebel and Besov classes;
3. the fact that recursive partitioning estimates achieve near-minimaxity over anisotropic smoothness classes.

From the CHA viewpoint, these results all follow from the fact that certain natural representations from harmonic analysis are the "optimal representation" for objects in the corresponding functional classes:

1. Fourier series are "optimal representations" for L^2-Sobolev classes.
2. Wavelets are "optimal representations" for L^p-Sobolev classes.
3. Anisotropic Haar bases are "optimal representations" for anisotropic smoothness classes.

The CHA viewpoint would say that results in mathematical statistics (near-minimaxity results for Fourier series, wavelets, etc.) are simple consequences of these more fundamental facts.

An appealing benefit of the CHA point of view is the fact that the central preoccupation of CHA is to rapidly compute such "optimal representations."

When the CHA viewpoint succeeds fully, one has therefore very practical methods. For example, since Fourier transforms, wavelet transforms and anisotropic transforms all have fast algorithms, the noise removal algorithms derived from CHA have fast algorithms—which also have near-optimality properties for appropriate classes $\mathscr{F}$.

On the other hand, one has to admit frankly that the CHA viewpoint is rather arrogant. It supposes that it is more interesting or important to get minimax or nearly minimax estimates by a certain point of view—through development of fast algorithms for newly invented decompositions—rather than to develop minimax or nearly minimax estimates in some other way.

1.3. *CHA and analysis of singularities.* Looking at the CHA viewpoint critically for a moment, one sees certain important open questions. Computational harmonic analysis has shown that Fourier series estimates, wavelets and straightforward recursive partitioning schemes are the optimal representations for certain classes $\mathscr{F}$. So each representation has its own specific domains of expertise and its own limitations. Despite their nice behavior in their respective domains of expertise, these representations do *not* provide simple nearly minimax estimates of objects with edges. For example, for a nice wavelet thresholding estimate $\hat{f}^{\text{wave}}$, and $\mathscr{F} = \text{HORIZ}^{\alpha}(C_1, C_{\alpha})$ with $1 < \alpha \leq 2$, we have the result that

$$\sup_{\mathscr{F}} \text{MSE}\left(\hat{f}^{\text{wave}}, f\right) \geq C \cdot n^{-1}, \qquad n \to \infty. \tag{1.4}$$

On the other hand, as we already know from, for example, [29] and [19],

$$M^*\left(n, \text{HORIZ}^{\alpha}(C_1, C_{\alpha})\right) = O\left(n^{-2\alpha/\alpha+1}\right). \tag{1.5}$$

So wavelets are not nearly minimax for objects with edges when $\alpha > 1$. More precisely, if we suppose that the edges in an image are C^{α} regular curves, $\alpha > 1$, then the minimax rate of convergence improves correspondingly as α improves; but the rate of convergence of wavelet estimators does not improve correspondingly: they suffer a speed limit of n^{-1}. Related speed limits apply to other schemes based on CHA ideas: sinusoids; dyadic recursive partitioning schemes.

In sum, the existing methods of CHA do not yield nearly minimax estimates on this simple edge model.

This is because CHA has not yet solved the problem of "analysis of singularities" in its most general form; CHA does not yet offer an algorithm which provides efficient representation of smooth objects with singularities along submanifolds of the ambient space (e.g., discontinuities along curves in image data, discontinuities along surfaces in three-dimensional data).

For an object with a discontinuity along a curve in dimension 2, there are $O(m)$ wavelet coefficients exceeding $1/m$ in amplitude. Indeed, at scale 2^{-j}, thee are $O(2^j)$ wavelets which "feel" the discontinuity and those have coefficients of size greater than or equal to 2^{-j}. As the performance of wavelet thresholding estimates is dependent on the number of big coefficients [12, 17,

21, 18, 13, 15] this limited decay rate of wavelet coefficients places a "speed limit" on the mean-squared error of wavelet estimates, which is recorded in (1.4) above. In order to achieve the bound (1.5) by thresholding in an orthogonal basis, we would need a much better basis, one which gave only $O(m^{2/(\alpha+1)})$ coefficients of size greater than or equal to $1/m$ for objects in $\text{HORIZ}^{\alpha}(C_1, C_{\alpha})$.

Unfortunately, the failure of harmonic analysis to address the problem of singularities is not well recognized in the computational harmonic analysis community (which helps explain why it is still open). It can be seen most clearly as a problem in the various cognate fields that CHA addresses—compression and denoising. Workers in compression realize that wavelet and Fourier compression methods do not represent edges efficiently. Workers in denoising realize that wavelet and Fourier methods do not efficiently remove noise from images with edges. Workers in CHA have not yet developed representations in their field which deal with edges efficiently.

1.4. *Wedgelets*. We are therefore interested in developing schemes of data representation which are general and can be applied to pixel data of more or less arbitrary type (i.e., not necessarily arising from the horizon model). Our representation will be based on objects we call *wedgelets*, which make up an overcomplete dictionary of atoms from which image data of arbitrary type may be synthesized. We will show that this representation is in principle well suited to recovering edges in the horizon model, by following a point of view which is natural for CHA. We will investigate three questions:

1. *Representation*—show that this is an "optimal representation" for objects in the horizon model.
2. *Algorithms*—develop a fast algorithm for obtaining an approximate representation of pixel-level datas, which may be used on clean or noisy data, arising from the horizon model or not.
3. *Estimation*—develop a way to invoke the fast algorithm so that when applied to noisy empirical data from the horizon model it achieves near-minimax behavior for recovering the underlying noiseless object.

We give solutions to all three problems. We show that wedgelets achieve nearly the minimax description length for objects in horizon classes. We develop a fast algorithm for obtaining atomic decompositions of noisy image data into wedgelets. The algorithm finds decompositions which correspond to special *edgelet-decorated* recursive partitions of the image; among such partitions, it optimizes complexity-penalized sum-of-squares. Owing to known oracle inequalities, this approach has certain near-ideal mean-squared error properties. Near-minimaxity over horizon classes follows directly.

The wedgelet approach can be successfully applied outside the horizon model. As an immediate generalization of our results, we show that the complexity-penalized estimator can recover, in a nearly minimax fashion, objects which are indicators of star-shaped sets having C^{α} boundaries.

From the CHA point of view, the contribution of this paper is the following. We point out an important general question—analysis of singularities—and we make contributions to CHA of singularities in a special case. We show that overcomplete systems of atoms, selected in an adaptive nonlinear fashion, can provide an acceptable analysis of singularities in a special case. This points to the possibility that simple, flexible harmonic analysis tools might be developed some day to deal with more realistic models of images.

2. Edgelets. We begin with some terminology and notation. A dyadic square S is the collection of points $\{(x_1, x_2)\colon [k_1/2^j, (k_1+1)/2^j] \times [k_2/2^j, (k_2+1)/2^j]\}$, where $0 \le k_1, k_2 < 2^j$ for an integer $j \ge 0$. For clarity, we will sometimes write $S(k_1, k_2, j)$, so that, for example, $S(0,0,0)$ is the unit square $[0,1]^2$, and so that, in the set ting of the data collection model of the introduction, if n is an n-by-n grid with $n = 2^J$ dyadic, then the individual pixels are the n^2 cells $S(k_1, k_2, J)$, $0 \le k_1, k_2 < n$.

Suppose we take vertices $v_1, v_2 \in [0,1]^2$ and consider the line segment $e = \overline{v_1 v_2}$. Following the literature of computer vision, we call such a segment an *edgel* (for edge element). If we consider only edgels connecting vertices $(k_1/n, k_2/n)$ at pixel corners. There are order $O(n^4)$ such edgels. For our purposes, we are seeking algorithms of order $O(n^2)$ or as near to that as we can get. The use of collections of cardinality $O(n^4)$ edgels will lead to unworkable algorithms, and so we seek a reduced-cardinality substitute.

Recalling that $n = 2^J$, take the collection of all dyadic squares at scales $0 \le j \le J$, that is, all dyadic squares down to pixel level, but not at finer levels. Fix a quantum of resolution $\delta = 2^{-J-K}$ for $K \ge 0$. On each dyadic square, traverse the boundary in clockwise fashion starting at the upper right corner and mark off equispaced vertices a distance δ apart. As δ is dyadic and it divides the perimeter length of every dyadic square with sidelength greater than or equal to $1/n$, there are precisely $M_j = 4 \cdot 2^K \cdot 2^{J-j}$ vertices marked out in this fashion on a dyadic square S with side $2^{-j} \ge 1/n$. Call this collection of vertices $V(S)$; label the vertices according to the order they are encountered in the clockwise boundary traverse, so that $V(S) = \{v_{i,S}\colon 0 \le i < M_j\}$. If we consider any two dyadic squares which have interesting boundaries, along the intersection of the boundaries the two squares have the same vertices in common; under our labelling system we might have $v_{i,S} = v_{i',S'}$ even though $i \ne i'$. For later use, we let $\mathscr{V}(n, \delta)$ denote the collection of all vertices in all $V(S)$ where S is dyadic of sidelength greater than or equal to $1/n$, and we let $\mathscr{L}(n)$ be latticework of all horizontal and vertical lines in the square at spacing $1/n$.

Within each dyadic square S, consider the collection of all edgels connecting vertices on the boundary of S:

$$E_\delta(S) = \left\{e = \overline{v_{i,S} v_{i',S}}\colon 0 \le i_1, i_2 < M_j\right\}.$$

There are $\binom{M_j}{2}$ such edgels in total.

DEFINITION 2.1. For given dyadic n and δ, the set of *edgelets* is the collection $\mathscr{E}_{n,\delta}$ of all edgels belonging to some $E_\delta(S)$ for some dyadic square S of sidelength 2^{-j}, $0 \le j \le J$.

We note that edgelets only connect vertices on the boundary of a dyadic square, so that although the family of edgelets is built from $O(n^2)$ vertices, it contains many fewer than $O(n^4)$ edgels. In fact, as $\binom{M_j}{2} \approx M_j^2/2$, we have

$$\begin{aligned}\#\mathscr{E}_{n,\delta} &= \sum_{j=0}^{J} \sum_{k_1,k_2=0}^{2^j-1} \#E_\delta(S(k_1,k_2,j)) = \sum_{j=0}^{J} 2^{2j} \binom{M_j}{2} \\ &\approx \sum_{j=0}^{J} 2^{2j} M_j^2/2 = \sum_{j=0}^{J} 2^{2j} \cdot 8 \cdot 4^K \cdot 2^{2(J-j)} \\ &= 8(\log_2(n) + 1)\delta^{-2}.\end{aligned}$$

For example, suppose that $\delta = 1/n$. Then there are just four vertices $v_{i,S}$ associated with any dyadic square with sidelength $1/n$; these are of course the corners of the squares, and we have

$$\#\mathscr{E}_{n,1/n} \approx 8 \cdot (\log_2(n) + 1) \cdot n^2.$$

Although there are order $O(n^2)$ pixels and order $O(n^4)$ edgels can be defined based on pixel corners, the collection of edgelets at that scale has a cardinality only logarithmically larger than $O(n^2)$. It follows that exhaustive searches through the collection of edgelets can run much faster than exhaustive searches through the collection of edgels.

Despite reduced cardinality, the dictionary of edgelets is expressive. It consists of edgels at a variety of scales, locations and orientations. A relatively small number of edgelets can be used as a substitute for any single edgel.

LEMMA 2.2. *Any edgel with endpoints anywhere in* $[0,1]^2$ *can be approximated within Hausdorff distance* $\delta/2 + 1/n$ *by a continuous chain* $(e_1, e_2, \ldots, e_m)$ *of edgelets* $e_i \in \mathscr{E}_{n,\delta}$, *where the number* m *of edgelets required is bounded by* $8\log_2(n)$ *for* $n > 2$.

We will prove this in Section 5. From the remark following the proof, we have:

COROLLARY 2.3. *Let* $H(t)$ *be a continuous "horizon" function,* $0 \le H(t) \le 1$ $\forall\ t \in [0,1]$. *Let* $\Gamma = \{(t, H(t))\}$ *be the associated horizon set in* $[0,1]^2$, *and suppose that this set can be approximated to within Hausdorff distance* ε *using at most* m *edgels with arbitrary vertices. Then this curve may be approximated within Hausdorff distance* $\varepsilon + \delta$ *using at most* $8\log_2(n) \cdot m$ *edgelets, for* $n > 2$, $m \ge 2$.

We will see below that the logarithmic factor can often be omitted, yielding a number of edgelets at most proportional to the underlying number of edgels.

3. Wedgelets. The dictionary of edgelets, while conceptually interesting, has little to do with approximation of image data of the type ordinarily obtained: that is, arrays $(y_{i_1,i_2})_{i_1,i_2=0}^{n-1}$ of pixel values. Edgelets are not functions and canot make up a basis for the space of numerical arrays. However, they do allow us to describe a convenient collection of such functions: the wedgelets.

Let S be a dyadic square; say that an edgelet $e \in E_\delta(S)$ is *nondegenerate* if it does not lie entirely on a common edge of S. A nondegenerate edgelet traverses the interior of S, and so splits S into two pieces, exactly one of which contains the segment of the boundary of S starting at $v_{0,S}$ and ending at $v_{1,S}$. Label the indicator of that piece $w_{e,S}$ and call this the *wedgelet* defined by e. Let

$$W_\delta(S) = \{1_S\} \cup \{w_{e,S}: e \in E_\delta(S) \text{ nondegenerate}\};$$

this collection of functions expresses all ways of splitting S into two pieces by edgelet splits, including the special case of not splitting at all.

DEFINITION 3.1. For given dyadic n and δ, $\mathscr{W}(n, \delta)$ is the collection of all wedgelets belonging to some $W_\delta(S)$, for some dyadic square S of sidelength 2^{-j}, $0 \le j \le J$.

3.1. *Optimal representation.* There is a sense in which wedgelets furnish near-optimal representations of objects in horizon classes. The following key estimate is proven in Section 5.3.

LEMMA 3.2. *Let $H(t)$ be a continuous horizon function, $0 \le H(t) \le 1$ $\forall$ $t \in [0,1]$. Suppose that $H \in \text{HÖLDER}^\alpha(C_\alpha)$, $1 \le \alpha \le 2$, and that $H \in \text{HÖLDER}^1(C_1)$ as well. Let $f(x_1, x_2) = 1_{\{x_2 \ge H(x_1)\}}$ be the associated "black-and-white image" defined on $[0,1]^2$. Suppose that $n > 2$ and let $2 \le m \le n$. There is a superposition of m' wedgelets,*

$$\hat{f}_{m'}(x_1, x_2) = \sum_{i=1}^{m'} \alpha_i w_i(x_1, x_2) \tag{3.1}$$

with

$$m' \le 8 \cdot (C_1 + 2) \cdot m$$

achieving an approximation error

$$\|f - \hat{f}_{m'}\|_2^2 \le K^\alpha \cdot C_\alpha \cdot m^{-\alpha} + \delta.$$

The object $\hat{f}_{m'}$ constructed in the proof of this lemma is itself a "black-and-white image," taking values 0 and 1 only. The expression (3.1) involves binary coefficients $\alpha_i \in \{0,1\}$. The wedgelets in the expression are selected

from a finite set of cardinality $\#\mathscr{W}(n, \delta)$. Hence each term in $\hat{f}_{m'}$ can be represented by a bit string of length $1 + \lceil \log_2(\#\mathscr{W}(n, \delta)) \rceil$ and so $\hat{f}_{m'}$ can be represented by a bit string of total length $\ell \le m' \cdot (1 + \lceil \log_2(\#\mathscr{W}(n, \delta)) \rceil)$.

For a desired error of approximation ε, pick a counting number $m(\varepsilon) \ge 2$ so that $K^\alpha \cdot m^{-\alpha} \le \varepsilon^2/2$. Pick $n(\varepsilon)$ and $\delta(\varepsilon)$ so that n is dyadic and $n \ge m$ and so that δ is dyadic and $\delta < \varepsilon^2/2$. Doing this in the obvious way gives functions $n(\varepsilon)$ and $\delta(\varepsilon)$ with $\log_2(\#\mathscr{W}(n(\varepsilon), \delta(\varepsilon))) \le \text{Const} \cdot \log_2(\varepsilon^{-1})$, for $0 < \varepsilon < 1$, where the constant depends on α, K^α and C. Lemma 3.2 constructs an object $\hat{f}_{m'}$ achieving error bounded by

$$\|f - \hat{f}_{m'}\|_{L^2[0,1]^2} \le \varepsilon, \tag{3.2}$$

and a corresponding description length

$$\ell(\varepsilon) \le \varepsilon^{-2/\alpha} \cdot \text{Const} \cdot \log_2(\varepsilon^{-1}), \ \varepsilon \to 0. \tag{3.3}$$

This description length is (within logarithmic factors) the best one can typically do for objects in $\text{HORIZ}^\alpha(C_1, C_\alpha)$. To explain what we mean, we quote from [14]. Let $\mathscr{F}$ be a compact set of functions in $L^2[0,1]^2$. Let ℓ be a fixed counting number and let $E_\ell \to \{0,1\}^\ell$ be a functional which assigns a bit string of length ℓ to each $f \in \mathscr{F}$. Let D_ℓ: $\{0,1\}^\ell \to L^2[0,1]^2$ be a mapping which assigns to each bit string of length ℓ a function. The coder–decoder pair (E_ℓ, D_ℓ) will be said to achieve distortion less than or equal to ε over $\mathscr{F}$ if

$$\sup_{f \in \mathscr{F}} \| D_\ell(E_\ell(f)) - f \|_{L^2[0,1]^2} \le \varepsilon. \tag{3.4}$$

We define the minimax description length as

$$L^*(\varepsilon, \mathscr{F}) = \min\{\ell : \exists\ (E_\ell, D_\ell) \text{ achieving distortion } \le \varepsilon \text{ over } \mathscr{F}\}. \tag{3.5}$$

This measures precisely the number of bits it is necessary to retain to be sure that the reconstruction of any $f \in \mathscr{F}$ will be accurate to within ε.

When $\mathscr{F}$ is not a finite set, then $L^*(\varepsilon, \mathscr{F}) \to \infty$ as $\varepsilon \to 0$, and the rate of this growth becomes of interest. In many interesting cases, $L^*(\varepsilon, \mathscr{F}) \asymp \varepsilon^{-1/\alpha}$ or $L^*(\varepsilon, \mathscr{F}) \asymp \varepsilon^{-1/\alpha} \log(1/\varepsilon)^\beta$ for some $\alpha, \beta > 0$. A crude measure of growth —insensitive to the difference between $\varepsilon^{-1/\alpha}$ and $\varepsilon^{-1/\alpha}\log(1/\varepsilon)^\beta$—is the *optimal exponent*

$$\alpha^*(\mathscr{F}) = \sup\{\alpha : L^*(\varepsilon, \mathscr{F}) = O(\varepsilon^{-1/\alpha}), \ \varepsilon \to 0\}. \tag{3.6}$$

From calculations in [19] we know that

$$\alpha^*(\text{HORIZ}^\alpha(C_1, C_\alpha)) = \alpha/2.$$

The bounds (3.2) and (3.3) say that wedgelet descriptions can achieve this optimal exponent.

3.2. *Atomic decomposition.* In our definition, we have taken the wedgelets w as functions of (x_1, x_2). We also find it useful to think of them as arrays of numbers. Given a wedgelet $w(x, y)$ we let $\tilde{w}(i_1, i_2)$ denote a pixel-level average of w, and let $\tilde{w}$ denote the array of all such averages.

In the spirit of much work in computational harmonic analysis, we think of the wedgelets as a collection of *atoms*, and we are interested in approximate atomic decompositions of arrays $y = (y(i_1, i_2))$ of the form

$$y = \sum_{w \in \mathcal{W}(n,\delta)} \alpha_w \tilde{w} + \text{Error}. \tag{3.7}$$

There is no unique way to approach this problem, because the collection $\mathcal{W}_{n,\delta}$ is over complete. Indeed it has cardinality larger than that of a basis. The vector space of n-by-n arrays has dimension n^2; but $\#\mathcal{W}(n, \delta) \geq 6 \cdot (J + 1) \cdot 4^K \cdot n^2$, which is larger than n^2 by a logarithmic factor. Following Mallat and Zhang [31], we call such a collection of elements a *dictionary* of atoms, with each one possessing a position, scale and (in some cases) pronounced orientation. This dictionary is *complete*, since it contains a subset which is a basis for the vector space of n-by-n arrays: the indicators of all pixels [as $1_S \in W(S)$ for every $S = S(i_1, i_2, J)$].

Using this basis, we can always obtain a trivial representation of the form (3.7): set $\alpha_w = 0$ unless $w = 1_S$ for some pixel $S = S(i_1, i_2, J)$, and then $\alpha_w = y(i_1, i_2)$. This always gives Error = 0. However, we are interested in representations which use only a small number of coefficients and yet have good approximation; this trivial representation uses n^2 coefficients in general.

Lemma 3.2 implies that, for objects arising from discretizing images with edges, there will exist good approximate representations with far fewer than n^2 wedgelets. How to find them? A variety of algorithms have been proposed for atomic decomposition. These can be divided into two groups:

1. *practically effective methods*, which run efficiently on current computers but cannot guarantee to find representations with near-optimal sparsity (matching pursuit [31], basis pursuit [6] and best-ortho-basis [8] are examples);
2. *theoretically effective methods*, which do guarantee to find good approximations, but require in principle enumeration of all subsets of a large collection of candidate decompositions [19]. Such "methods" cannot be used on large-scale problems.

In special cases, there are specific algorithms which run rapidly and which give near-best results for objects in certain classes; see [15] for an example. Our goal in this paper is to develop an algorithm of this form.

3.3. *Wedgelet analysis*. In a setting where we have a dictionary $\mathcal{D} = \{\phi\}$ of atoms, there are two tasks which we can distinguish:

1. analysis—compute inner products of dictionary elements with data, obtaining $\langle \phi, y \rangle$, for all $\phi \in \mathcal{D}$;
2. synthesis—givena subset of the dictionary (ϕ_i) and coefficients (a_i) form $y = \sum a_i \phi_i$.

These two items are in principle different, although they are deeply linked.

What we were studying up to now was decomposition, and this involves both elements of analysis and synthesis. Analysis is necessary to identify atoms which should appear in the decomposition; synthesis is necessary to actually construct the approximate representation. Analysis is an important ingredient of any practical algorithm for atomic decomposition, and the ability to rapidly analyze is a prerequisite for the ability to rapidly decompose and synthesize. There is an explicit discussion of the importance of this in [6]; but one can say that standard methods of atomic decomposition (matching pursuit [31], basis pursuit [6] and best-ortho basis [8]) depend for their practicality on the ability to rapidly compute all the inner products between data to be decomposed and the dictionary.

Now for an arbitrary collection of inner products $\langle \phi, y \rangle$, for all $\phi \in \mathscr{D}$, there is no hope for fast calculation. Naively, to compute one single inner product would take as many n^2 calculations; as there are at least $n^2 \log_2(n)$ elements in $\mathscr{W}(n, \delta)$, one is led to a figure approaching $O(n^4 \log(n))$ complexity. For problems where n may be in the hundreds or thousands this order of complexity seems daunting.

A basic reason for defining the wedgelet dictionary as we have done is the prospect of computational efficiency. Let us formalize our discussion:

DEFINITION 3.3. *The wedgelet analysis* (*WA*) *of an array y is the vector of all inner products*

$$\langle \tilde{w}, y \rangle \qquad \forall\, w \in \mathscr{W}(n, \delta); \tag{3.8}$$

this vector has $N(n, \delta) = \#\mathscr{W}(n, \delta)$ *entries.*

We anticipate that it is possible to rapidly compute an approximate WA, and the developments in this paper are based on the following:

MAJOR PREMISE. *It is possible to calculate an approximate wedgelet analysis in no more than* $C \cdot N \log(N)$ *flops, where the constant C reflects the degree of approximation.*

A discussion of how to perform such a computation rapidly, and of the size of the approximation error and so on, is properly a topic in computational harmonic analysis and lies outside the scope of the present article. However, we will take the premise as given for the purposes of this paper. We plan to describve algorithms for rapid wedgelet analyusis elsewhere.

4. Recursive partitioning. To obtain an effective algorithm, we now develop notions of adaptive partitioning.

We begin with inheritance terminology. We say that four adjacent dyadic squares are *siblings* if their union is a dyadic square, which we call their *parent*. The four siblings are called *children* of their common parent. The operation of partitioning a parent square into its four children is called a *standard quad-split*. The operation of combining four siblings into one parent

is called a *standard quad-merge*. Given a dyadic square S, its *ancestors* are its parent, grandparent, ...; its *descendants* are its children, grandchildren ... .

DEFINITION 4.1. A *recursive dyadic partition* (RDP) is any partition of $[0,1]^2$ reachable by applying the following production rules recursively:

(i) The trivial partition $\mathscr{P} = \{[0,1]^2\}$ is an RDP.

(ii) If $\mathscr{P} = \{S_1, \ldots, S_i, \ldots, S_m\}$ is an existing RDP, then the partition obtained by applying a standard quad-split on one of the squares in $\mathscr{P}$ is another RDP.

We let RDP(n) denote the collection of RDPs where all squares are of sidelength greater than or equal to $1/n$.

There are many types of RDPs. The uniform partition of depth j consists of all 4^j dyadic squares of sidelength 2^{-j}; it is spatially homogeneous. There are also spatially inhomogeneous partitions, such as the partition of depth $\log_2(n)$ which contains at levels $j = 1, 2, \ldots, \log_2(n)$ the square $S(k_1, k_2, j)$, $(k_1, k_2) \in \{(1,0),(0,1),(1,1)\}$, and which also contains $S(0,0,\log_2(n))$. This partition is very fine in a small neighborhood of $(0,0)$ and gets increasingly coarse as one moves away from $(0,0)$.

The definition of RDP is based on refining existing RDPs. There is also a coarsening operation. An RDP $\mathscr{P}$ has the form $\mathscr{P} = \{S_1, S_2, \ldots, S_m\}$ for some m and some dyadic squares S_i. Such a partition must contain four squares which are siblings. If we apply a standard quad-merge to those four siblings, we get a new, coarser RDP.

Each RDP is associated with a quadtree Q rooted at $[0,1]^2$ whose terminal nodes are the S_i, and whose interior nodes are the ancestors of the S_i. We can think of the quadtree as a road map describing how to produce the given RDP by applying standard quad-splits starting from the unit interval $[0,1]^2$. Also, by applying standard quad-merges to an RDP, one can create various coarser RDPs; each such coarser RDP corresponds to its own quadtree, which is a subtree of the quadtree corresponding to the original RDP. In a very natural sense refining an RDP is the same as "growing" the associated tree and coarsening an RDP is the same as "pruning" the associated tree.

DEFINITION 4.2. An *edgelet-decorated recursive dyadic partition* is an RDP $\mathscr{P}$ in which each member S can (optionally) be decorated with (at most) one nondegenerate edgelet in $E_\delta(S)$. Such a decorated partition $\mathscr{P}$ induces a subordinate partition $\overline{\mathscr{P}}$ in which decorated squares of $\mathscr{P}$ are split into two pieces along the edgelet boundary. Let ED-RDP(n, δ) be the collection of all such partitions.

In an RDP, all splits are parallel to the axes and are located halfway along the sides, and each split always results in four children. In an ED-RDP there is the additional possibility of "splitting-diagonally" some of the squares into

two pieces, but we are forbidden to do this recursively—squares which have been split into two arbitrary pieces rather than four dyadic squares may not be further subdivided. The terminal regions of an ED-RDP are either dyadic squares or else polygons; we use the letter P for either type of piece. Also, we often use the typographical device of an overline to indicate whether the object in question is an RDP (e.g., $\mathscr{P}$) or the partition derived from an ED-RDP (e.g., $\overline{\mathscr{P}}$).

A key fact about ED-RDPs is that they can provide good approximations to objects in class $\text{HORIZ}^{\alpha}(C_1, C_{\alpha})$.

Let $f = f(x_1, x_2)$ be a function in $L^2[0,1]^2$. Let $\overline{\mathscr{P}}$ be an ED-RDP. Define $\text{Ave}\{f|\overline{\mathscr{P}}\}$ in the obvious way, as the function $\bar{f}(x_1, x_2)$ which is constant on each piece P of $\overline{\mathscr{P}}$:

$$\bar{f}(x_1, x_2) = \sum_{P \in \overline{\mathscr{P}}} \bar{f}_P 1_P(x_1, x_2),$$

where of course $\bar{f}_P = \text{Ave}\{f(x_1, x_2): (x_1, x_2) \in \mathscr{P}\}$. Section 5.4 proves:

LEMMA 4.3. *Suppose that* $H \in \text{HÖLDER}^{\alpha}(C_{\alpha})$, $1 \le \alpha \le 2$, *and that* $H \in \text{HÖLDER}^{1}(C_1)$ *as well. Let* $n > 2$ *and* $2 \le m \le n$. *Let* $f(x_1, x_2) = 1_{\{x_2 \ge H(x_1)\}}$ *be the associated "black-and-white image" defined on* $[0,1]^2$. *There exists an ED-RDP with fewer than* $m' = 8 \cdot (C_1 + 2) \cdot m$ *elements having an approximation error* $K^{\alpha} C_{\alpha} m^{-\alpha} + \delta$. *More precisely*,

$$\inf_{\#\mathscr{P} \le m'} \sup_{f \in \mathscr{F}} \left\| f - \text{Ave}\left\{f|\overline{\mathscr{P}}\right\} \right\|_{L^2}^2 \le K^{\alpha} \cdot C_{\alpha} \cdot m^{-\alpha} + \delta,$$

where K^{α} *depends on* α *only*.

This lemma shows that approximation schemes based on RDPs achieve bounds similar to what can be achieved with wedgelets. In fact there is a close connection. The approximant $\bar{f} = \text{Ave}\{f|\overline{\mathscr{P}}\}$ is a kind of atomic decomposition using wedgelets. Indeed, it may be written as a sum

$$\bar{f} = \sum_{w} a_w w.$$

This is due to the fact that, by definition, $\mathscr{W}(n, \delta)$ contains every indicator 1_S of every dyadic square with sidelength greater than or equal to $1/n$, while, for any region P arising by splitting a dyadic square S into two pieces along a nondegenerate edgelet e, we have

$$1_P = a_P \cdot 1_S + b_P \cdot w_e.$$

Indeed, either P or $S \setminus P$ must be the wedge $\text{supp}(w_e)$, and so either $a_P = 0$ and $b_P = 1$ or else $a_P = 1$ and $b_P = -1$.

Notice, however, that because $\bar{f}$ arises from adaptive recursive partitioning, it is not built in an arbitrary fashion from wedgelets. Various constraints are enforced by the recursive partitioning which mean that only certain wedgelets can appear simultaneously in this expansion. In our case, these

constraints will be seen to be useful, because they make possible a fast algorithm; but they do not seem to lead to significantly worse upper bounds on approximate error.

5. Proofs of key lemmas. Now that we have established RDP concepts and terminology, we can easily prove the lemmas stated in Sections 2–4. We pause to do this before proceeding.

5.1. *Preparation*

LEMMA 5.1. *Let $\mathscr{S}$ be a finite collection of dyadic squares. If the squares in the collection have disjoint interiors, there exists on RDP with those squares as members. If all the squares are of sidelength greater than or equal to $1/n$, there exists an RDP(n) with those squares as terminal squares.*

PROOF. Consider the following procedure. Starting from $\mathscr{P} = \{[0,1]^2\}$, recursively apply the following rule: if any member of $\mathscr{S}$ is a proper subset of a square S' in the current partition, apply a standard quad-split to S', obtaining a new current partition.

For any specific $S \in \mathscr{S}$, the procedure must eventually arrive at a partition which includes S, because it will always decide to split any square properly containing S. Once it arrives at a partition that includes such an $S \in \mathscr{S}$, it will not split S because (by the non-overlapping interiors condition), no members of $\mathscr{S}$ lie inside S. Hence, the procedure terminates at a partition in which every square in $\mathscr{S}$ appears as an element. □

LEMMA 5.2. *The coarsest RDP containing a set $\mathscr{S}$ of squares as members is precisely the RDP whose quadtree contains all children of all ancestors of all squares in $\mathscr{S}$. The algorithm of Lemma* 5.1 *produces this RDP.*

If $S = [0,1]^2$, so that it strictly speaking has no ancestors, we still regard S itself as a child of ancestors.

PROOF OF LEMMA 5.2. Consider a quadtree associated with a partition in which all members of $\mathscr{S}$ appear also as members of the partition. If S' is an ancestor of an $S \in \mathscr{S}$, then S' must be an interior node of the quadtree. But then the quadtree must contain all the children of S'. In short any such quadtree must contain, either as interior or terminal nodes, all the children of all ancestors of all intervals $S \in \mathscr{S}$. The smallest such quadtree is exactly the tree that contains only the ancestors and children of ancestors, and it corresponds to an RDP with elements of $\mathscr{S}$ as members.

The algorithm of Lemma 5.1 splits only nodes which are ancestors of members of $\mathscr{S}$. Hence it produces an RDP whose quadtree contains only the children of ancestors of members of $\mathscr{S}$.

LEMMA 5.3. *A square S of sidelength greater than or equal to $1/n$ yields at most $4\log_2(n) + 1$ ancestors and children of ancestors.*

PROOF. Count them. □

LEMMA 5.4. *Let $H \in \text{HÖLDER}^1(C)$, and let $\Gamma = \{(t, H(t))\}$. Then for $\mathscr{S}_j$ the collection of all dyadic squares of sidelength 2^{-j} having nonempty intersection with Γ,*

$$\#\mathscr{S}_j \le (C + 2) \cdot 2^j.$$

PROOF. Let $t_{k,j} = k/2^j$. We simply interpret the Lipschitz condition

$$\sup_{t \in [t_{k,j}, t_{k+1,j}]} \left| H(t) - H(t_{k,j}) \right| \le C \cdot 2^{-j}$$

as saying that the curve Γ traverses at most $C + 2$ dyadic squares of side 2^{-j} as t runs from $t_{k,j}$ to $t_{k+1,j}$. □

5.2. *Proof of Lemma* 2.2. Let e be the given arbitrary edgel. We assume it is not purely vertical or purely horizontal and give the proof in that case only.

The edgel e intersects with the $(n + 1)$-by-$(n + 1)$ latticework $\mathscr{L}(n)$ of equispaced vertical and horizontal lines at spacing $1/n$. Let $\tilde{e}$ be the subset of e extending from the intersection point v_0 closest to one endpoint of e to the intersection point v_1 closest to the other endpoint of e.

There are vertices v_{i_0,S_0}, v_{i_1,S_1} belonging to our discrete set $\mathscr{V}(n, \delta)$ within a distance $\delta/2$ of the endpoints of $\tilde{e}$. Those dyadic points belong to dyadic squares S_0 and S_1, say. Either we may take these squares to be disjoint, or if this is impossible because these dyadic points belong to the same pixel, we take them to be identical. Set $\mathscr{S} = \{S_0, S_1\}$. Run Lemma 5.1 to construct the coarsest RDP $\mathscr{P}$ containing elements of $\mathscr{S}$.

Take the edgel $\tilde{e}$ from v_0 to v_1. For simplicity of exposition, speak of this edgel as if it goes "from" v_0 "to" v_1. We say that an edgel *traverses* a dyadic square *completely* if it both enters and leaves the square, that is, it intersects the square and does not terminate inside the square. Mark all the elements of the partition $\mathscr{P}$ which the edgel $\tilde{e}$ traverses completely. Let $\mathscr{S}^*$ be the collection of all marked squares. In each marked S, choose the vertices $v_{i_0,S}$ and $v_{i_1,S}$ in $V(S)$ which are nearest the points of entry and exit. Note that in neighboring marked squares S, S', the vertex closest to the exit point of S will be also the vertex closest to the entry point into S', since squares with boundaries in common have vertices in common.

There is a possibility of ties, where two vertices in $V(S)$ are equidistant from an entry or exit point. This should be resolved as follows. Since squares with boundaries in common have vertices in common, if there is a tie in the assignment in S [say, two vertices in $V(S)$ are equally close to the entry point of $\tilde{e}$ in S], there will a marked square adjacent to S, S' in which there is also a corresponding tie [say, two vertices in $V(S')$ are equally close to the exit

point of $\tilde{e}$]. The pairs of vertices involved in each tie, although assuming different labels in each case, are actually the same pairs of points in $[0,1]^2$. We can resolve the tie in any way we like (coin tossing?) but we should resolve it the same way in each of the two dyadic squares.

The selection of vertices $v_{i_0,S}$ and $v_{i_1,S}$ for each marked S gives us a list of edgelets $\{e(S)\colon S \in \mathscr{S}^*\}$ indexed by marked S. Because adjacent marked squares have entry points and exit points in common, this list of edgelets makes a chain—a connected set.

This chain never deviates from $\tilde{e}$ by more than $\delta/2$ in Hausdorff distance. Indeed, the intersection of $\tilde{e}$ with each marked S makes an edgel contained in S whose endpoints are at most $\delta/2$ away from the endpoints of the corresponding $e(S)$.

The edgel $\tilde{e}$ never deviates from e by more than $1/n$.

Hence, the chain never deviates from e by more than $1/n + \delta/2$.

It remains to count the number of members of the chain. This is no greater than the number $\#\mathscr{S}^*$ of marked squares. All marked squares must be children of ancestors of S_0 or S_1; S_0 has sat most $4\log_2(n) + 1$ children of ancestors. Similarly for S_1. Hence there are not more than $8\log_2(n) + 2$ edgelets in the chain. We can refine this slightly. If $n > 2$, and S_0 and S_1 have sidelength less than $1/2$, note that S_0 and S_1 always have in common five elements among their ancestors and children of ancestors: the unit square and its four children. The estimate $8\log_2(n) + 2$ double counts these five squares. In that case $m \le 8\log_2(n)$ will also work. If either S_0 or S_1 is of sidelength greater than $1/4$ one sees that $m \le 8\log_2(n)$ continues to work (for $n > 2$).

REMARK. Suppose that the endpoints of the edgel e lie in the $(n+1)$-by-$(n+1)$ latticework $\mathscr{L}(n)$ of equispaced vertical and horizontal lines at spacing $1/n$. Then the chain of edgelets comes within distance δ of e (rather than $\delta/2 + 1/n$).

Indeed, the edgel $\tilde{e}$ constructed in the proof can then be taken identical to e. The endpoints belong to the latticework, and these will be at most $\delta/2$ distance away from members of $\mathscr{V}(n,\delta)$, the endpoints of e, rather than $1/n + \delta/2$ away.

5.3. *Proof of Lemma* 3.2. Since the underlying horizon $H \in \text{HÖLDER}^{\alpha}(C_{\alpha})$, the polygonal approximation H_m obtained by linearly interpolating the $m+1$ points $(i/m, H(i/m))$ has error

$$\|H_m - H\|_{\infty} \le K^{\alpha} \cdot C_{\alpha} \cdot m^{-\alpha}, \qquad m = 1, 2, \ldots, \tag{5.1}$$

for K a constant depending only on α.

Second, we note that it is enough to establish the result for dyadic m, that is, m of the form 2^j for $1 \le j \le J$. The result for general m in the range $2 \le m \le n$ will follow with a possibly larger constant. So let m be dyadic.

View the linear interpolant H_m as a chain of m edgels. Let $v_{i,0}$ and $v_{i,1}$ be the endpoints of the ith edgel. These have x_1-coordinates which are integer multiples of $1/m$, and, by the dyadicity assumption on m, these therefore belong to the latticework $\mathscr{L}(m)$ at scale $1/n$. Let $\tilde{v}_{i,0}$ and $\tilde{v}_{i,1}$ be points with the same x_1-coordinates as the original $v_{i,0}$ and $v_{i,1}$ but belonging to our discrete set of vertices $\mathscr{V}(n, \delta) \subset \mathscr{L}(m)$; they are chosen by the condition that they are closest members of $\mathscr{V}(m, \delta)$. (There is the possibility of ties; by following a consistent tie-breaking procedure, we can arrange that $\tilde{v}_{i+1,0} = \tilde{v}_{i,1}$.) These new points are at most distance $\delta/2$ from the original points. The chain of edgels $\tilde{v}_{i,0}\tilde{v}_{i,1}$ comes within a distance $\delta/2$ of the original chain.

As in the proof of Lemma 5.4, each edgel in the chain traverses at most $C_1 + 2$ dyadic squares of sidelength $1/m$. It can be approximated within distance $\delta/2$ by a chain of at most $(C_1 + 2)$ edgelets, each one associated with a different square, the squares being vertically adjacent.

The resulting chain of chains of edgelets is continuous as a whole, contains at most $(C_1 + 2) \cdot m$ edgelets and yields a function $H_{m,\delta}$ which obeys

$$\|H_m - H_{m,\delta}\|_\infty \le \delta.$$

Third, consider now the function

$$f_{m,\delta}(x_1, x_2) = 1_{\{x_2 \ge H_{m,\delta}(x_1)\}}.$$

We have

$$\|f - f_{m,\delta}\|_{L^2}^2 \le \|H - H_{m,\delta}\|_{L^1[0,1]} \le K^\alpha \cdot C_\alpha \cdot m^{-\alpha} + \delta, \tag{5.2}$$

so that $f_{m,\delta}$ has the degree of approximation we are seeking. We claim that $f_{m,\delta}$ is a superposition of m' wedgelets.

Take each edgelet associated with $H_{m,\delta}$. That edgelet is itself associated with a unique dyadic square S. Hence $H_{m,\delta}$ gives us immediately a corresponding finite collection $\mathscr{S}_{m,\delta}$ of dyadic squares. These dyadic squares all have common sidelength $1/m$. Dyadic squares of the same size are either identical or disjoint. The collection $\mathscr{S}_{m,\delta}$ of dyadic squares constructed in the building of the chain of chains therefore consists of squares with disjoint interiors.

The algorithm of Lemma 5.1 gives an RDP $\mathscr{P}_{m,\delta}$ which has all the squares in $\mathscr{S}_{m,\delta}$ as members. Let $\overline{\mathscr{P}}_{m,\delta}$ be the corresponding ED-RDP resulting from decorating the partition $\mathscr{P}_{m,\delta}$ so that squares $S \in \mathscr{S}_{m,\delta}$ are decorated with the corresponding edgelets.

We can count the cardinality of $\mathscr{P}_{m,\delta}$ as follows. It consists of all the ancestors and children of ancestors of squares in $\mathscr{S}_{m,\delta}$. Now any ancestor of a square in $\mathscr{S}_{m,\delta}$ has nonempty intersection with the curve $(t, H(t))$. We can bound the number of such ancestors using Lemma 5.4; the number of ancestors of sidelength 2^{-j} is at most $2^j \cdot (C_1 + 2)$. The number of children of ancestors of sidelength 2^{-j} is at most $4 \cdot 2^{j-1} \cdot (C_1 + 2)$. The total across all scales of the number of ancestors and children of ancestors can be

estimated as

$$4 \cdot \sum_{2^j \le m} 2^j \cdot (C_1 + 2) \le 8 \cdot (C_1 + 2) \cdot m.$$

This is a bound on the cardinality of $\mathscr{P}_{m,\delta}$.

We now turn to $\overline{\mathscr{P}}_{m,\delta}$. There are at most $(C_1 + 2) \cdot m$ squares in $\mathscr{S}_{m,\delta}$ and each one might potentially be decorated. Each decoration of a square increases the cardinality of $\overline{\mathscr{P}}_{m,\delta}$ over $\mathscr{P}_{m,\delta}$. Hence $\#\overline{\mathscr{P}}_{m,\delta} \le 8 \cdot (C_1 + 2) \cdot m$. We have proven the lemma. □

5.4. *Proof of Lemma* 4.3. Lemma 3.2 showed how to construct a partition $\overline{\mathscr{P}}_{m,\delta}$ and a function $f_{m,\delta}$ that is constant on each piece of the partition $\overline{\mathscr{P}}_{m,\delta}$. In the technical sense, it is $\overline{\mathscr{P}}_{m,\delta}$-measurable, by standard properties of conditional expectation, if $\mathscr{P}$ is an arbitrary partition, and f' is $\mathscr{P}$-measurable, then

$$\| f - \mathrm{Ave}\{ f | \mathscr{P} \} \|_{L^2} \le \| f - f' \|_{L^2}.$$

Applying that in this case

$$\left\| f - \mathrm{Ave}\left\{ f | \overline{\mathscr{P}}_{m,\delta} \right\} \right\|_{L^2} \le \| f - f_{m,\delta} \|_{L^2}.$$

Hence $\overline{\mathscr{P}}_{m,\delta}$ gives us an RDP achieving the indicated error bounds. The same complexity bounds on the partition that were developed for Lemma 3.2 apply here. □

6. Fast recursive partitioning. We are now in a position to define a specific principle for processing noisy data (1.1). Our goal is to find a partition with low cardinality which fits the data well. The approach we use is of exactly the same type as employed in [18, 19, 15].

Let $y(i_1, i_2)$ be an array of pixel-level data. Suppose we are given an ED-RDP $\overline{\mathscr{P}}$. In the vector space of n-by-n arrays, consider the vector subspace $L(\overline{\mathscr{P}})$ of all arrays arising from linear combinations $\sum_{P \in \overline{\mathscr{P}}} C_P \tilde{1}_P$, where $\tilde{1}_P$ is the array of pixel averages of the function 1_P. Define $\widetilde{\mathrm{Ave}}\{ y | \overline{\mathscr{P}} \}$ as the array resulting from least-squares projection of y onto $L(\overline{\mathscr{P}})$.

We fix a *complexity penalty factor* λ (to be discussed later) and we define the *complexity-penalized sum of squares*

$$\mathrm{CPRSS}(\overline{\mathscr{P}}, \lambda) = \left\| y - \widetilde{\mathrm{Ave}}\left\{ y | \overline{\mathscr{P}} \right\} \right\|^2 + \lambda^2 \# \overline{\mathscr{P}};$$

this associates with an ED-RDP a figure of merit which combines both the residual sum of squares $\| y - \widetilde{\mathrm{Ave}}\{ y | \overline{\mathscr{P}} \} \|^2$ and the complexity of the partition. The empirically *optimal* ED-RDP is then

$$\overline{\mathscr{P}}^* = \operatorname*{argmin}_{\text{ED-RDP}(n,\delta)} \mathrm{CPRSS}(\overline{\mathscr{P}}, \lambda),$$

and our empirical estimate is

$$\hat{f}^* = \widetilde{\mathrm{Ave}}\left\{ y | \overline{\mathscr{P}}^* \right\}.$$

This estimate is obtained by selecting an optimal ED-RDP according to the CPRSS criterion. This is a kind of empirical atomic decomposition using discrete wedgelets. Indeed, it may be written as

$$\hat{f}^* = \sum_w a_w \tilde{w},$$

for the same reasons as (3.7). Because $\hat{f}^*$ arises from adaptive recursive partitioning, it is not built in an arbitrary fashion from wedgelets. Various constraints are enforced by the recursive partitioning which mean that only certain wedgelets can appear simultaneously in this expansion. In our case, these constraints are useful, because they make possible a fast algorithm to find $\hat{f}^*$.

Recall our major premise of Section 3: that it is possible to rapidly obtain an approximate wedgelet analysis. The following result uses the major premise as a starting point. It will follow from this result and the major premise that one can essentially obtain $\hat{f}^*$ in $O(N \log_2^2(N))$ time, where $N = \#\mathscr{W}(n, \delta)$.

THEOREM 6.1. *Suppose we are given the wedgelet analysis of y—the collection of all inner products $(\langle \tilde{w}, y \rangle : w \in \mathscr{W}(n, \delta))$. There is an algorithm for finding $\hat{f}^*$ starting from this collection which operates in order $O(N)$ time, where $N = \#\mathscr{W}(n, \delta)$.*

The algorithm is based on ideas of dynamic programming and backward induction; it is similar in basic outline to the "best-ortho-basis" algorithm of Coifman and Wickerhauser [8] and to the optimal tree pruning algorithm in the CART book [4]; see also [2, 15, 37]. In the remainder of this section, we first describe a basic decomposability property of the CPRSS, then give an algorithm for minimizing CPRSS. Finally we give the proof of Theorem 6.1.

6.1. *Additivity of CPRSS.* The objective function has certain additivities which imply that it can be optimized sequentially. For a dyadic square S, let $\overline{\mathscr{P}}[S]$ denote an ED-RDP of S, let $y[S]$ denote the subarray of our original pixel array consisting only of pixels belonging to S, and let $\mathrm{CPRSS}(\overline{\mathscr{P}}[S], \lambda; S)$ denote the *localized* complexity penalized sum-of-squares

$$\mathrm{CPRSS}\big(\overline{\mathscr{P}}[S], \lambda; S\big) = \Big\| y[S] - \widetilde{\mathrm{Ave}}\big\{y[S] | \overline{\mathscr{P}}[S]\big\} \Big\|^2_{\ell^2[S]} + \lambda^2 \#\overline{\mathscr{P}}[S],$$

where $\ell^2[S]$ denotes the ℓ^2-norm over the indicated subarrays.

The localized CPRSS has two useful additivity properties:

1. *Global decomposition*—suppose that $\mathscr{P}$ is an RDP (not an ED-RDP) and that $\overline{\mathscr{P}}$ is a finer ED-RDP. For a dyadic square S in $\mathscr{P}$, let $\overline{\mathscr{P}}[S]$ be partition of S induced by restriction of $\overline{\mathscr{P}}$ to S. Then we have the identity

$$\mathrm{CPRSS}(\overline{\mathscr{P}}, \lambda) = \sum_{S \in \mathscr{P}} \mathrm{CPRSS}\big(\overline{\mathscr{P}}[S], \lambda; S\big),$$

expressing the global CPRSS in terms of analogous localized quantities.

2. *Local decomposition*—again let S be a dyadic square, let $\bar{\mathscr{P}}[S]$ denote a partition of S and let $\mathscr{P}[S]$ denote any coarser partition of S. In a notation borrowed from the computer language C, for $S' \subset S$, let $\bar{\mathscr{P}}[S][S']$ denote the restriction of the partition $\bar{\mathscr{P}}[S]$ to induce a partition of S':

$$\text{CPRSS}(\bar{\mathscr{P}}[S], \lambda; S) = \sum_{S' \in \mathscr{P}[S]} \text{CPRSS}(\bar{\mathscr{P}}[S][S'], \lambda; S'). \tag{6.1}$$

Now let $\text{CPRSS}^*(\lambda; S)$ denote the optimal value of the localized CPRSS:

$$\min_{\bar{\mathscr{P}}[S]} \text{CPRSS}(\bar{\mathscr{P}}[S], \lambda; S).$$

By (6.1) this new quantity obeys two inheritance relations, which we can state as follows. Let $\mathscr{P}(S, e)$ denote the two-element partition of S created by a split along edgelet $e \in \mathscr{E}_\delta(S)$. Let $\{S\}$ be the trivial partition of S into one set:

1. *Inheritance at coarse scales*—suppose that S has sidelength greater than $1/n$, and let S_i, $i = 1, 2, 3, 4$, denote the four children of S. Then

$$\text{CPRSS}^*(\lambda; S) = \min \begin{cases} \text{CPRSS}(\{S\}, \lambda; S), \\ \min_{e \in \mathscr{E}(S)} \text{CPRSS}(\mathscr{P}(S, e), \lambda; S), \\ \sum_{i=1}^{4} \text{CPRSS}^*(\lambda; S_i). \end{cases} \tag{6.2}$$

2. *Inheritance at fine scales*—on the other hand, if S has sidelength $1/n$ so that it has no children which can belong to any RDP(n), then

$$\text{CPRSS}^*(\lambda; S) = \min \begin{cases} \text{CPRSS}(\{S\}, \lambda; S), \\ \min_{e \in \mathscr{E}(S)} \text{CPRSS}(\mathscr{P}(S, e), \lambda; S). \end{cases} \tag{6.3}$$

6.2. *Tree pruning algorithm.* The inheritance relations (6.2) and (6.3) lead to a hierarchically-organized algorithm for minimizing CPRSS.

ALGORITHM (Optimal decorated quadtree). This algorithm finds the ED-RDP achieving CPRSS* by breadth-first, bottom-up pruning. When it terminates, it holds a quadtree whose terminal nodes are the squares of the RDP $\mathscr{P}$ associated with the optimal RDP and whose labels indicate the decorations, if any, attached to those squares in the optimal ED-RDP.

- Build the complete quadtree of depth $\log_2(n)$.
- Initialize: Label each node S with
  - $a_S = \text{CPRSS}(\{S\}, \lambda; S)$
  - $b_S = \min_{e \in \mathscr{E}(S)} \text{CPRSS}(\mathscr{P}(S, e), \lambda; S)$,
  - $e_S =$ the e achieving b_S.
  - Set level $j = \log_2(n)$.
  - Loop: for each S of sidelength 2^{-j},

- Inherit: If $j < \log_2(n)$ set $d_S = \sum_i c_{S_i}$ (sum over children of S); while if $j = \log_2(n)$ set $d_S = +\infty$ (there are no children).
- Tournament: Compute $c_S = \min(a_S, b_S, d_S)$.
  - If the minimum is achieved by a_S, mark node S "Terminal: undecorated"
  - If the minimum is achieved by b_S, mark node S "Terminal: decorated by edgelet e_S."
  - If the minimum is achieved by d_S, mark node S "Interior."
- Prune: If S is marked "Terminal," prune away from the current quadtree the four children of S and the subtrees (if any) rooted at those children.
- Set $j = j - 1$;
- If $j \geq 0$ goto Loop.

6.3. *Proof of Theorem* 6.1. The algorithm is fast.

The dominant computational burden is in step Initialize. We need, to begin with, quantities measuring the size and energy of y on dyadic squares

$$\overline{y^2}[S] = \frac{1}{\#S} \sum_S y^2(i_1, i_2), \qquad \bar{y}[S] = \frac{1}{\#S} \sum_S y(i_1, i_2).$$

The collection of all these numbers, for all squares S of sidelength greater than or equal to $1/n$, can be computed from the collection of all $\langle \tilde{1}_S, y \rangle$ (which we have assumed given) in order $O(n^2)$ time. The formulas for a_S and b_S are just

$$a_S = \#S \cdot \left(\overline{y^2}[S] - \bar{y}[S]^2 \right) + \lambda$$

and

$$b_S = \#S \cdot \overline{y^2}[S] - \langle \tilde{w}, y \rangle^2 / \langle \tilde{w}, \tilde{w} \rangle^2 + 2 \cdot \lambda,$$

where $\tilde{w} = \tilde{w}_{e(S),S}$. Hence, all these quantities can be computed in order $O(\#\mathscr{W}(n, \delta))$ operations.

In the main loop of the algorithm, we process each node of the complete quadtree of depth $\log_2(n)$. At each node we have order $O(1)$ computations to do. There are $\frac{4}{3} \cdot n^2$ nodes of this quadtree. This takes $O(n^2)$ work in general.

In summary, we can initialize and then traverse the whole tree in order $O(\#\mathscr{W}(n, \delta)) + O(n^2) = O(\#\mathscr{W}(n, \delta))$ time.

The algorithm correctly computes the optimal CPRSS because it is just a systematic application of the inheritance relations (6.2) and (6.3).

7. Main result. In addition to rapid computation, the estimator $\hat{f}^*$ can have, when the complexity penalty is chosen appropriately, near-minimax mean-squared error, simultaneously over a broad collection of horizon classes.

THEOREM 7.1. *Let* $\delta = \delta_n = 2^{-\lceil J\beta \rceil}$, *where* $\beta > 4/3$. *Base the wedgelet dictionary and analysis on* $\mathscr{W}(n, \delta_n)$. *Fix* $\xi > 8$, *and set*

$$\lambda = \left(\xi \cdot \sigma \cdot \left(1 + \sqrt{2 \log_e(\#W(n, \delta_n))} \right) \right)^2. \tag{7.1}$$

Let $\hat{f}^*$ *denote the complexity-penalized estimator produced with this* λ. *For this estimator*

$$\sup_{\mathscr{F}} \mathrm{MSE}\left(\hat{f}^*, f\right) \le O(\log(n)) \cdot M^*(n, \mathscr{F}), \tag{7.2}$$

where $\mathscr{F} = \mathrm{HORIZ}^{\alpha}(C_1, C_{\alpha})$ *for some* $\alpha \in [1, 2]$.

This is true, whatever $\alpha \in [1, 2]$ may be, with a simple choice of λ; it is not necessary to adapt λ to the unknown f. Comparing this result with (1.5) and the discussion surrounding it, we see that the method improves on traditional means of harmonic analysis.

In the next subsections we prove this result, according to the following outline. In Section 7.1, we describe how the MSE of the estimator $\hat{f}^*$ may be controlled in terms of the ideal MSE one would suffer if an oracle revealed the ideal partitioning. In Section 7.2, we calculate the ideal MSE. In Section 7.3 we combine the pieces into a proof.

7.1. *Oracle inequalities*. 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 [17, 18, 19] call the *ideal MSE*

$$\mathscr{M}^*\left(\hat{\Phi}, f\right) = \inf\left\{\mathrm{MSE}\left(\hat{f}, f\right): \hat{f} \in \hat{\Phi}\right\}.$$

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 $\mathscr{M}^*(\hat{\Phi}, f)$ as a target, and seek true estimators which can approach this target. We do not expect to do as well with a true estimator as one can do with an oracle. But it turns out that in a range of examples [17–19, 15], one can find estimators which do achieve this to within logarithmic factors. The inequalities which establish this are called *oracle inequalities*, because they compare the risk of valid procedures with the risk achievable by idealized procedures which depend on oracles. Compare related ideas of Birgé and Massart [3] and of Foster and George [23]. A simple corollary of results in [17–19, 15] gives an oracle inequality for the estimator $\hat{f}^*$ of this paper.

For each $\overline{\mathscr{P}} \in \text{ED-RDP}(m, \delta)$, consider the fixed partition estimator $\hat{f}(\cdot; \overline{\mathscr{P}}) = \widetilde{\mathrm{Ave}}\{y|\overline{\mathscr{P}}\}$. The family of such estimators is $\hat{\Phi} = \{\hat{f}(\cdot; \mathscr{P}): \overline{\mathscr{P}} \in \text{ED-RDP}(n, \delta)\}$. For obvious reasons, we call $\mathscr{M}^*(\hat{\Phi}, f)$ also $\mathscr{M}^*(\text{IDEAL PARTITIONING}, f)$, as it represents the risk we would suffer by estimating using the ideal ED-RDP, if we only knew to use it.

THEOREM 7.2. *For the minimum CPRSS estimator* $\hat{f}^*$ *defined with* λ *in* (7.1).

$$\mathrm{MSE}(\hat{f}^*, f) \le \mathrm{Const} \cdot \log(\#W(n, \delta)) \times \left(\frac{\sigma^2}{n^2} + \mathscr{M}^*(\text{IDEAL PARTITIONING}, f) \right) \qquad \forall f. \tag{7.3}$$

In short, empirical adaptive partitioning (with an appropriate penalization) comes within log-factors of the performance of ideal adaptive partitioning.

7.2. *Ideal risk calculation.* Consider the linear model

$$y_{i_1, i_2} = \delta_{i_1, i_2} + \sum_{j=1}^{p} \beta_j x^{(j)}_{i_1, i_2} + z_{i_1, i_2}, \qquad 0 \le i_1, i_2 < n.$$

Here the $(x^{(j)}_{i_1, i_2})$, $j = 1, \dots, p$, are predictor arrays, the β_j are prediction coefficients, and (δ_{i_1, i_2}) is an (nonrandom) array orthogonal to the linear span of the $x^{(j)}$'s. The z_{i_1, i_2} make a Gaussian white noise with variance σ^2. Suppose we estimate the coefficients $(\beta_j)_{j=1}^p$ by a least squares fit of the form

$$\min_{\beta} \left\| y - \sum_{j=1}^{p} \beta_j x^{(j)} \right\|_{\ell^2}^2,$$

producing $(\hat{\beta}_j)$ and predicted values $\hat{y}_{i_1, i_2} = \sum_{j=1}^p \hat{\beta}_j x^{(j)}_{i_1, i_2}$. The predictive mean-squared error is then

$$E\|\hat{y} - y\|_{\ell^2}^2 = \|\delta\|_{\ell^2}^2 + p\sigma^2, \tag{7.4}$$

that is, a traditional Bias^2 + Variance expression.

Now the operator $\widetilde{\mathrm{Ave}}\{y|\overline{\mathscr{P}}\}$ which we have defined in Section 6 is a least-squares projection operator of the form just described. Indeed, set $p = \#\overline{\mathscr{P}}$, and enumerate the pieces of the partition as $\overline{\mathscr{P}} = \{P_j : j = 1, \dots, p\}$. Then set $x^{(j)} = \tilde{1}_{P_j}$. With this correspondence the predicted values $\hat{y} = \widetilde{\mathrm{Ave}}\{y|\overline{\mathscr{P}}\}$ precisely, and $\delta = \tilde{f} - \widetilde{\mathrm{Ave}}\{\tilde{f}|\overline{\mathscr{P}}\}$. We conclude that

$$E\left\|\widetilde{\mathrm{Ave}}\{y|\overline{\mathscr{P}}\} - y\right\|_{\ell^2}^2 = \left\|\tilde{f} - \widetilde{\mathrm{Ave}}\{\tilde{f}|\overline{\mathscr{P}}\}\right\|_{\ell^2}^2 + \#\overline{\mathscr{P}} \cdot \sigma^2. \tag{7.5}$$

Now let $P_n f$ be the operator that replaces a function $f(x_1, x_2)$ by its array of pixel-level averages $\tilde{f}(i_1, i_2)$. Then of course $n^{-1}\|P_n f\|_{\ell^2} \le \|f\|_{L^2[0,1]^2}$. Let $f_{m, \delta}$ be the function guaranteed by Lemma 4.3, and as usual $\tilde{f}_{m, \delta} = P_n f_{m, \delta}$. Then

$$\begin{aligned} n^{-1} \cdot \left\| \tilde{f} - \widetilde{\mathrm{Ave}}\{\tilde{f}|\overline{\mathscr{P}}\} \right\|_{\ell^2} &\le n^{-1} \cdot \|\tilde{f} - \tilde{f}_{m, \delta}\|_{\ell^2} \\ &= n^{-1} \cdot \|P_n(f - f_{m, \delta})\|_{\ell^2} \\ &\le \|f - f_{m, \delta}\|_{L^2[0,1]^2}. \end{aligned}$$

Combining this with (7.5) we get

$$\mathrm{MSE}\left(\widetilde{\mathrm{Ave}}\{y|\bar{\mathscr{P}}\}, f\right) \le \|f - f_{m,\delta}\|^2_{L^2[0,1]^2} + \frac{\sigma^2}{n^2} \cdot \#\bar{\mathscr{P}}.$$

Now for $f \in \mathrm{HORIZ}^{\alpha}(C_1, C_{\alpha})$, Lemma 4.3 gives

$$\mathrm{MSE}\left(\widetilde{\mathrm{Ave}}\{y|\bar{\mathscr{P}}\}, f\right) \le K^{\alpha} \cdot C_{\alpha} \cdot m^{-\alpha} + \delta_n + \frac{\sigma^2}{n^2} \cdot m',$$

with $m' \le 8(C_1 + 2)m$. Define $\mu(\varepsilon, m) = K^{\alpha} \cdot C_{\alpha} \cdot m^{-\alpha} + \varepsilon^2 \cdot m$. Then this last display says that $\mathrm{MSE}(\widetilde{\mathrm{Ave}}\{y|\bar{\mathscr{P}}\}, f) \le \mu(\varepsilon, m)$, where $\varepsilon^2 = 8 \cdot \sigma^2(C_1 + 2)/n^2$. The ideal risk is then not bigger than what we can get by optimizing this expression in m over the range $2 \le m \le n$. Now in the range $1 \le \alpha \le 2$, we get that, for all sufficiently small $\varepsilon > 0$, the minimum obeys

$$\min_{2 \le m \le n} \mu(\varepsilon, m) \le (\varepsilon^2)^{\alpha/\alpha+1} \cdot (C_{\alpha})^{1/\alpha+1} \cdot K'_{\alpha} + \varepsilon^2.$$

Hence, for $\mathscr{F} = \mathrm{HORIZ}^{\alpha}(C_1, C_{\alpha})$, the ideal risk obeys

$$\begin{aligned} &\sup_{\mathscr{F}} \mathscr{M}^*(\text{IDEAL PARTITIONING}, f) \\ &\qquad \le \delta_n + K''_{\alpha} \cdot (C_{\alpha})^{1/\alpha+1} \cdot \left(\sigma^2/n^2\right)^{\alpha/(\alpha+1)}. \end{aligned} \tag{7.6}$$

7.3. *Completion of the proof.* We know from (1.5) that

$$M^*(n, \mathscr{F}) \ge c \cdot C_{\alpha}^{1/1+\alpha} \cdot n^{-2\alpha/1+\alpha}.$$

By assumption $\alpha \le 2$, and so $\delta_n = O(n^{-\beta}) = o(n^{-4/3}) = o(n^{-2\alpha/1+\alpha})$. Hence, from (7.6),

$$\sup_{\mathscr{F}} \mathscr{M}^*(\text{IDEAL PARTITIONING}, f) \le O(1) \cdot M^*(n, \mathscr{F}), \qquad n \to \infty;$$

the risk of ideal partitioning is within constant factors of the minimax risk. But from the oracle inequality (7.3) we know that the risk of empirical partitioning is within log factors of the risk of ideal partitioning, and so

$$\begin{aligned} \mathrm{MSE}(\hat{f}^*, f) &\le \mathrm{Const} \cdot \log(n) \cdot \left(\frac{\sigma^2}{n^2} + \mathscr{M}^*(\text{IDEAL PARTITIONING}, f)\right) \\ &\le O(\log(n)) \cdot M^*(n, \mathscr{F}), \qquad n \to \infty. \end{aligned}$$

8. Generalization. We briefly discuss two avenues of generalizations of the above results.

8.1. *Inhomogeneous boundaries.* In the paper so far, we have considered the case where the horizon function obeys certain Hölder conditions. Such conditions are of a very spatially homogeneous type—they impose the same type of condition on H in the vicinity of each point. We may consider instead functional classes $\mathrm{HORIZ}^{\alpha}_{p,q}(C_1, C_{\alpha})$ defined by the condition that the horizon function H belong to Besov ball $B^{\alpha}_{p,q}(C_{\alpha})$. Here $p, q > 0$ are scalars; α is a

smoothness index. The scale of Besov spaces (see, e.g., [33, 24]) includes various Hölder-type spaces, as well as L^2-Sobolev spaces; $B^{\alpha}_{\infty,\infty}(C)$ is very nearly $\text{HÖLDER}^{\alpha}(C)$, and so what we have been calling $\text{HORIZ}^{\alpha}(C_1, C_\alpha)$ is very nearly $\text{HORIZ}^{\alpha}_{\infty,\infty}(C_1, C_\alpha)$.

Horizon functions in certain Besov classes are spatially inhomogeneous, smooth in most of the domain, with potentially exceptional behavior on a special subset. An example is the case $\alpha = 2$, $p = 1$, $q = \infty$, which consists more or less of functions which are primitives of functions of bounded variation. Such functions behave in many respects like twice-differentiable functions at "most" points, but they are not necessarily even continuously differentiable.

Our main result extends easily to cover this more general scale of examples.

THEOREM 8.1. *Let $q \in (0, \infty]$ and $1 \le p \le \infty$. Let $\mathscr{F} = \text{HORIZ}^{\alpha}_{p,q}(C_1, C_\alpha)$ for $1 \le \alpha \le 2$. Then* (7.2) *holds for this $\mathscr{F}$.*

The heart of the matter is on display in (5.2). What we really need are bounds of the form

$$\|H - H_m\|_{L^1[0,1]} \le K^{\alpha} \cdot C_\alpha \cdot m^{-\alpha}, \qquad m = 2, 4, 8, \ldots,$$

where H_m is an approximant of H based on m edgels. So far in this paper we have controlled L^∞ approximation; but the weaker L^1 condition is all we really needed. The following lemma is based on well-known properties of Besov spaces (see [11]).

LEMMA 8.2. *Let $H \in B^{\alpha}_{p,q}(C)$, $0 < \alpha \le 2$, $p \ge 1$. The equispaced knot linear spline interpolant H_m obeys*

$$\|H - H_m\|_{L^1[0,1]} \le K^{\alpha}_{p,q} \cdot C \cdot m^{-\alpha}, \qquad m = 2, 4, 8, \ldots, \tag{8.1}$$

where $K^{\alpha}_{p,q}$ depends on (α, p, q) only.

Applying this lemma will immediately give a more general version of Lemma 3.2; the rest of the arguments in the paper go through verbatim, giving Theorem 8.1.

8.2. *General shapes*. There is nothing about the estimator $\hat{f}^*$ of (7.1) which specifically requires that the underlying object be of the "horizon" form. It may be applied to any data of the form (1.1); it makes no explicit assumptions about the type of object being estimates.

To illustrate this point, we now consider star-shaped objects. A star-shaped set $B \subset [0,1]^2$ has an origin $b_0 \in [0,1]^2$ from which every point of B is "visible," that is, such that the line segment $\{(1-t)b_0 + tb\colon\ t \in [0,1]\} \subset B$ whenever $b \in B$. We will show that the estimator $\hat{f}^*$ works well on objects $f = 1_B$. To do so, we define $\text{STAR-SET}^{\alpha}(C)$, a class of star-shaped sets with regular boundaries using a kind of polar coordinate system. (This type of class is studied at greater length in [19].) Let $\rho(\theta)\colon [0, 2\pi) \to [0,1]$ be a

radius function and $b_0 = (x_{1,0}, x_{2,0})$ be an origin with respect to which the set of interest is star-shaped. Define $\Delta_1(x) = x_1 - x_{1,0}$ and $\Delta_2(x) = x_2 - x_{2,0}$; then define functions $\theta(x_1, x_2)$ and $r(x_1, x_2)$ by

$$\theta = \arctan(-\Delta_2/\Delta_1), \qquad r = \left((\Delta_1)^2 + (\Delta_2)^2\right)^{1/2}.$$

For a star-shaped set, we have $(x_1, x_2) \in B$ iff $0 \le r \le \rho(\theta)$. In particular, the boundary ∂B is given by the curve

$$\beta(\theta) = (\rho(\theta)\cos(\theta) + x_{1,0}, \rho(\theta)\sin(\theta) + x_{2,0}). \tag{8.2}$$

The class STAR-SET$^\alpha(C)$ of interest to us can now be defined by

$$\text{STAR-SET}^\alpha(C) = \Big\{B: B \subset \left[\tfrac{1}{10}, \tfrac{9}{10}\right]^2, \tfrac{1}{10} \le \rho(\theta) \le \tfrac{1}{2}, \\ \theta \in [0, 2\pi), \rho \in \text{HÖLDER}^\alpha(C)\Big\}.$$

NOTE 1. Some star-shaped sets have more than one possible choice of origin b_0; different choices lead to different radius functions ρ; we demand only that *some* valid choice of b_0 lead to a ρ obeying the above conditions.

NOTE 2. We consider only the range $1 < \alpha \le 2$; $\alpha \le 1$ is excluded.

The actual objects of interest are the indicators of sets in STAR-SET$^\alpha$, so we introduce the functional class

$$\text{STAR}^\alpha(C) = \{f = 1_B: B \in \text{STAR-SET}^\alpha(C)\}. \tag{8.3}$$

THEOREM 8.3. *Let $\hat{f}^*$ denote the complexity-penalized estimator as described in Theorem* 7.1. *Let $\mathscr{F} = \text{STAR}^\alpha(C)$ for some choice $1 < \alpha \le 2$, $0 < C < \infty$. Then* (7.2) *holds for this $\mathscr{F}$.*

8.2.1. *Outline of proof.* The proof can be given an architecture paralleling closely the structure used in proving Theorem 7.1. We know from the risk lower bounds in, for example, [19], for a $c > \phi$ that

$$M^*(n, \mathscr{F}) \ge c \cdot C_\alpha^{1/1+\alpha} \cdot n^{-2\alpha/1+\alpha};$$

so if we can show that for $\mathscr{F} = \text{STAR}^\alpha(C)$, the ideal risk obeys

$$\sup_{\mathscr{F}} \mathscr{M}^*(\text{IDEAL PARTITIONING}, f) \le \delta_n + K^\alpha \cdot (C_\alpha)^{1/(\alpha+1)} \cdot \left(\sigma^2/n^2\right)^{\alpha/(\alpha+1)}.$$

then it will follow from $\delta_n = o(n^{-2\alpha/1+\alpha})$ that the ideal risk obeys

$$\sup_{\mathscr{F}} \mathscr{M}^*(\text{IDEAL PARTITIONING}, f) \le O(1) \cdot M^*(n, \mathscr{F}), \qquad n \to \infty;$$

the risk of ideal partitioning is within constant factors of the minimax risk. It will then follow from the oracle inequality that over each class STAR$^\alpha(C)$ the risk of $\hat{f}^*$ is within a log factor of minimax.

The required estimate for ideal risk can be stated as follows:

LEMMA 8.4. *Let* $\mathscr{F} = \text{STAR}^\alpha(C)$, $1 < \alpha \le 2$, $0 < C < \infty$. *For each* $f \in \mathscr{F}$ *there exists a corresponding ED-RDP with fewer than* $m' = K_1 \cdot m + K_2$ *elements having an approximation error* $O(m^{-\alpha}) + \delta$. *More precisely*,

$$\inf_{\#\mathscr{P} \le m'} \sup_{f \in \mathscr{F}} \left\| f - \text{Ave}\{f|\overline{\mathscr{P}}\} \right\|_{L^2}^2 \le K^\alpha \cdot C_\alpha \cdot m^{-\alpha} + \delta,$$

where K^α *depends on* α *only*.

The proof of Lemma 8.4 proceeds in the same fashion as the analogous fact for the horizon case, Lemma 4.3. There are three stages, the first establishing an approximation to the boundary of B by general edgels, the second obtaining a subsidiary approximation to $f = 1_B$ by edgelets and the third counting the number of wedgelets involved. The first stage is as follows.

LEMMA 8.5 (Edgel approximation). *Let* $B \in \text{STAR-SET}^\alpha(C)$. *A unit-speed parametrization of the boundary of B exists*; *call it* $\beta(s)$. *Below we describe an* "*espalier construction*" *which yields*, *for each dyadic m obeying* $m \ge 2^{\mu_0}$, *a polygonal approximation* β_m *with these properties*:

(i) *All the edgels belonging to* β_m *have endpoints in* $\mathscr{L}(m) \cap \partial B$.
(ii) *Each edgel belonging to* β_m *has length* $\le K_1/m$, K_1 *a fixed constant*.
(iii) *At least* 50% *of the edgels belonging to* β_m *have length greater than or equal to* $1/m$.
(iv) *Any dyadic square S of side* $1/m$ *which intersects the interior of an edgel in* $\{\beta_m\}$ *intersects one and only one such edgel*.

The proof will be given further below, after the espalier construction is explained. The lemma implies that

$$\|\beta_m - \beta\|_\infty \le K^\alpha \cdot C \cdot m^{-\alpha}, \qquad m = 2^{\mu_0}, 2^{\mu_0+1}, \ldots, \tag{8.4}$$

by (ii) above, the fact that β is Hölderian, and the fact that linear interpolation of Hölderian functions obeys estimates based on the length h of the longest line segment; here, of course $h = K_1/m$. Compare (5.1) and (8.1), where a similar principle was invoked. This lemma also implies that the list E_m of edgels in β_m has cardinality

$$\#E_m \le 2 \cdot |\partial B| \cdot m, \tag{8.5}$$

by (iii) above, and the obvious fact that in comparing arclengths we have Length(β_m) $\le$ Length(β), owing to Euclid.

LEMMA 8.6 (Edgelet approximation). *Starting from* E_m, *we construct a list* $\overline{E}_m$ *of edgelets and a corresponding curve* $\overline{\beta}_m$ *by approximating each edgel in*

β_m by a corresponding edgelet. We can do this in a way which guarantees the following properties:

(i) *The dyadic squares of scale* $1/m$ *which intersect* β_m *also intersect* $\overline{\beta}_m$.
(ii) *Any dyadic square of side* $1/m$ *which intersects* $\overline{\beta}_m$ *intersects the interior of one and only one edgelet* $e \in \overline{E}_m$.
(iii) $\| \overline{\beta}_m - \beta_m \|_\infty \le \delta$.
(iv) *The number of edgelets obeys the bound*

$$\#\overline{E}_m \le C_1 \cdot K_1 \cdot m \cdot |\partial B| + C_2 .$$

Now of course we have

$$\| \overline{\beta}_m - \beta \|_\infty \le K^\alpha \cdot C \cdot m^{-\alpha} + \delta , \tag{8.6}$$

by (iii) and (8.4). To finish up, we need to estimate the number of pieces in an ED-RDP generating a set with $\overline{\beta}_m$ for a boundary curve. By Lemma 5.2, we know this reduces to counting the ancestors of the dyadic squares associated with the edgelets in $\overline{E}_m$.

LEMMA 8.7 (Counting ancestors). *Starting from the list of* $\overline{E}_m$ *of edgelets we can construct an ED-RDP* $\overline{P}_m$ *decorated by these edgelets with cardinality bounded by*

$$\#\overline{P}_m \le K_1 \cdot m \cdot |\partial B| + K_2 , \tag{8.7}$$

and the approximation error specified in Lemma 8.4.

It remains to describe the espalier construction and to prove these three lemmas.

8.2.2. *Espalier construction*. An *espalier* is a gardener's device—a uniform latticework to which one can attach shoots of a growing vine to manage the shape of the plant. Here we use the latticework $\mathscr{L}(m)$ to "tie down" our curve β to certain points on the latticework, and approximate the curve by line segments in between the places where it is tied down. By judicious choice of where to tie the curve down we will obtain a decomposition of the curve into Lipschitz graphs, each of which behaves as in the horizon model.

The espalier construction has these stages:

1. *Initialization*—take $\Gamma_1 = \mathscr{L}(m) \cap \partial B$.
2. *Pruning*—Γ_1 is in general an uncountably infinite set, since some pieces of ∂B may coincide perfectly with a vertical or horizontal segment of the latticework. We now "prune" this set by replacing any compoent of Γ_1 which is a line segment by its endpoints together with points having coordinates which are both integer multiples of $1/m$. An at-most countable set Γ_2 results.
3. *Labelling*—we now label each member γ of this discrete set by labels $\ell_1(\gamma)$ and $\ell_2(\gamma)$:
   3a. *Orientation*—$\ell_1(\gamma) \in \{V, H, B\}$ depending on the type of line segment the curve is intersecting (intersection with **V**ertical, **H**orizontal, **B**oth).

3b. *Quality*—$\ell_2(\gamma) \in \{E, G, P\}$ (**E**xcellent, **G**ood, **P**oor) depending on the angle between the tangent to the curve and the segment it is intersecting: if the angle is $< \pi/6$, mark "P"; if the angle is $\in [\pi/6, \pi/3]$ mark "G"; if the angle is $> \pi/3$, mark "E". If γ is precisely at a lattice crossing (i.e., $\ell_1(\gamma) = B$), then choose the label based on the "better" of the two possible labels. In such an event, for later use, modify the corresponding label $\ell_1(\gamma)$ to reflect the orientation of the "better" of the two crossings.

4. *Chaining*—now gather a subset of the points $\gamma \in \Gamma_2$ into chains which are consistent locally with a function $y = f(x)$ or a function $x = f(y)$:
   4a. The points in Γ_2 can be ordered circularly by the numerical value of the angle the tangent makes with the x_1-axis, so there is a clear notion of predecessor and successor.
   4b. JAt least one point $\gamma \in \Gamma_2$ is marked E. Initiate a chain labelled from $\{V, H\}$ according as the label $\ell_1(\gamma)$.
   4c. Grow the chain as far it is possible to do so without starting a "bad link." Suppose that γ is the most recently added point in the chain. Starting at the immediately succeeding point, iterate through successor points γ', looking for the first occurrence of a point γ' where either (i) $\ell_1(\gamma) = \ell_1(\gamma')$ (here we are continuing the existing chain, and we add to it γ) or (ii) $\ell_1(\gamma) \neq \ell_1(\gamma')$ but $\ell_2(\gamma) = E$ [here we are ending the old chain and starting a new one, with a new orientation; we mark the old chain ended at γ; we mark γ' as the first member of a new chain labelled from $\{V, H\}$ according as $\ell_1(\gamma')$].
   4d. Continue systematically until returning to the starting point. Deal with the termination in the obvious way, merging the last chain with the first if they have the same $\{V, H\}$ labelling.

The result of this construction is a finite collection of finitely many chains, every vertex in a chain having an identical ℓ_1-labelling. The points of such a chain, once connected together by line segments, generate a Lipschitz graph of the form either $y = f(x)$ or $x = f(y)$. The maximum slope of such a chain is bounded by an absolute constant depending on the choice of constants $\cos(\pi/3)$ in the labelling and on the constants (α, C) underlying ∂B. Each chain consists of "good links" having length greater than or equal to $1/m$ and less than or equal to K_1/m.

Our description of the construction makes a number of assertions (such as "there exists a point labelled 'E'"; "chains have links less than or equal to K_1/m," etc.) which need to be supported. Underlying this justification is the following lemma.

LEMMA 8.8. *Let $\beta(\theta)$ be the boundary curve of a $B \in$ STAR-SET$^\alpha(C)$, $1 < \alpha \leq 2$, as in* (8.8). *Then β has a continuous tangent vector field and a unique unit-speed parametrization $\beta(s)$. This unit-speed cuve is uniformly continuous, uniformly in $B \in$ STAR-SET$^\alpha(C)$. Indeed, there exists a uniform modulus of continuity $\omega(\delta; \alpha, C)$ satisfying $\omega(\delta) \to 0$ as $\delta \to 0$ and so that,*

for every β *arising from a unit-speed parametrization of such a boundary curve* $B \in \text{STAR-SET}^{\alpha}(C)$,

$$\left\| \dot{\beta}(s) - \dot{\beta}(s+\delta) \right\| \le \omega(\delta), \qquad \delta > 0.$$

In effect, this just follows from the uniform modulus of continuity of $\beta(\theta)$.

LEMMA 8.9. (1) *Each point in* Γ_1 *can be given a well-defined angle of intersection with the lattice* $\mathscr{L}(m)$.

(2) *There is* $\mu_0 = \mu_0(\alpha, C)$ *so that for* $m \ge 2^{\mu_0}$:

(2a) *If* β *enters a vertical column by crossing a vertical line of* $\mathscr{L}(m)$ *at angle greater than* $\pi/6$, *it continues across the column and exits the column by crossing the vertical line of* $\mathscr{L}(m)$ *on the opposite side of the same column.*

(2b) *In moving from one side of the column to the other*, β *intersects at most* K_1 *dyadic squares of side* $1/m$.

(2c) *If* β *reenters the column after leaving*, *it will not*, *during its second passage through the column*, *intersect any of the same dyadic squares it has already intersected.*

Of course similar statements hold with occurrences of "vertical" replaced by "horizontal."

(3) *There exists a point* $\gamma \in \Gamma_2$ *labelled* $\ell_2(\gamma) = E$.

(4) *Each chain contains at least two vertices and the link that joins them.*

(5) *Each chain stops the first time a point* γ' *marked* $\ell_2(\gamma') = P$ *is accepted into the chain.*

PROOF. For the reader's convenience, we recall the points to be proved, in italic text:

(1) *Each point in* Γ_1 *can be given a well-defined angle of intersection with the lattice* $\mathscr{L}(m)$. Since $\dot{\beta}(s)$ is continuous there is a well-defined tangent to β at each point of Γ_1, so (after fixing a sense of orientation) we can unambiguously define the "angle" between $\dot{\beta}$ and any fixed direction. Typically, a point of Γ_1 is at the crossing of β with a horizontal line or a vertical line in $\mathscr{L}(m)$, but not both. In either of these typical cases, the angle of crossing is unambiguously defined.

If the intersection point is precisely at the crossing of β both with a horizontal and a vertical line, we ignore the crossing which is "most nearly tangent" [i.e., smallest value of |sin(angle of intersection)|, flipping a coin in case of ties], and record the angle with the other crossing.

(2a) *If* β *enters a vertical column by crossing a vertical line of* $\mathscr{L}(m)$ *at angle greater than* $\pi/6$, *it continues across the column and exits the column by crossing the vertical line of* $\mathscr{L}(m)$ *on the opposite side of the same column.* Since $\dot{\beta}$ is uniformly continuous, there is a $\delta = \delta(\varepsilon) > 0$ so with $\dot{\beta}(s) = (\dot{h}(s), \dot{v}(s))$,

$$|s - s_0| < \delta \quad \Rightarrow \quad \left|\dot{h}(s) - \dot{h}(s_0)\right| < \varepsilon.$$

Picking $\varepsilon_0 = \cos(\pi/3)/2$, we get that for $\delta_0 = \delta(\varepsilon_0)$ that if $|\dot{h}(s_0)| > \cos(\pi/3)$ then $|s - s_0| < \delta_0$ implies

$$|h(s) - h(s_0)| > |\dot{h}(s_0)||s - s_0|/2. \tag{8.8}$$

Hence if $|\dot{h}(s_0)| > \cos(\pi/3)$, the curve β continues in a monotone fashion in the x_1-direction throughout the interval $s_0 < s < s_0 + \delta_0$. Taking $2^{-\mu_0} < \delta_0 \cdot \varepsilon_0$, we get that, for $m = 2^{\mu}$, $\mu \geq \mu_0$, there is an s_1 in the range $s_0 < s_1 < s_0 + \delta_0$ with

$$|h(s_1) - h(s_0)| = \frac{1}{m}; \tag{8.9}$$

that is, the curve exits the vertical column bounded on one side by $x_1 = h(s_0)$ and the other by $x_1 = h(s_1)$.

(2b) *In moving from one side of the column to the other*, β *intersects at most* K_1 *dyadic squares of side* $1/m$. Indeed, as $|v(s_1) - v(s_0)| < |s_1 - s_0|$ we get, by (8.8) and (8.9),

$$|v(s_1) - v(s_0)| \leq 2/\cos(\pi/3) \cdot (1/m) = A/m, \quad \text{say.}$$

The curve can therefore only intersect squares in this vertical column within vertical distance A/m of the point of entry; there are at most $K_1 = A + 2$ such squares of side $1/m$.

(2c) *If* β *reenters the column after leaving*, *it will not*, *during its second passage through the column*, *intersect any of the same dyadic squares it has already intersected*. For $\dot{\beta}$ to reenter a vertical column after leaving it, $\dot{\beta}(s)$ will have to be strictly vertical at some point s_1 in between exit and reentry.

It will leave at a point γ' marked $\ell_2(\gamma') = P$. Indeed, had the point beeen marked "E" or "G," then in the step marked 4, chaining, the espalier construction would have selected to *continue* the chain rather than to *break* the chain, as it actually did. See (2a).

On the other hand, the preceding point γ in the chain will have been marked $\ell_2(\gamma) = \mathrm{G}$ or E. Recall the argument above at (8.8) and (8.9). Let s_0 be the value of the arclength parameter yielding $\gamma = \beta(s_0)$, and let s_1 be the value yielding $\gamma' = \beta(s_1)$. For the same value δ_0 used in that argument, from the fact that $|\dot{h}(s_0)| > \cos(\pi/3)$, we know that

$$|\dot{h}(s)| > 0, \qquad s_1 < s < s_1 + \delta_0.$$

In short, even after the curve moves out of the column, it continues in the "away" direction at least for another arclength δ_0 units. But now the same modulus of continuity argument applies to $v(s)$ that was used for $h(s)$, and as $\dot{v}^2 + \dot{h}^2 = 1$, we have

$$|\dot{v}(s)| > \cos(\pi/3), \qquad s_1 < s < s_1 + \delta_0. \tag{8.10}$$

Exactly the same manipulations as at (8.8) and (8.9) show that, for some s_2 in the range $s_1 < s_2 < s_1 + \delta_0$,

$$|v(s_2) - v(s_1)| = \frac{1}{m}.$$

In short, long before $\dot{h}$ can vanish, the curve will have crossed a *horizontal* line.

At this point, the tangent vector will be pointing into (say) the upper right quadrant of the plane. But at this point, the previously encountered square is in (say) the lower left quadrant anchored at the current point. Hence, in order to reenter the previous square in the column it is necessary for the tangent to turn through at least 90°.

At this point we invoke star-shapedness. The square in question subtends —relative to the origin b_0—an angle of size at most $\Delta\theta_m$, since by hypothesis $\rho > 0.1$ and $1/m < 0.05$ (say). A star-shaped curve cannot reenter a sector it has left, until it goes a full circle. So if the curve $\beta(\theta)$ cannot reenter the square within an angle distance $\Delta\theta_m$, star-shapedness forbids it ever to do so (short of reentry after a full circle). The modulus of continuity shows that for sufficiently large μ_0, and all $m \geq 2^{\mu_0}$, the tangent cannot turn by 90° in an angular distance $\Delta\theta_m$.

(3) *There exists a point* $\gamma \in \Gamma_2$ *labelled* $\ell_2(\gamma) = E$. As β is a closed curve, the image of $\dot{\beta}(s)$ covers the circle. Hence there exists a point s_1 at which $\dot{\beta}(s_1)$ is exactly vertical.

The argument for (2a) gives us a δ_0 so that, for $s \in [s_1 - \delta_0, s_1 + \delta_0]$, $|\dot{v}(s)| > \sin(\pi/3)$. Because $2^{-\mu_0} < \delta_0 \cdot \varepsilon_0 \cdot \varepsilon_0$ the curve must cross in this interval a horizontal line from $\mathscr{L}(m)$. This intersection point will belong to Γ_2 and will be labelled $\ell_2 = E$.

(4) *A chain contains at least two points and the link that contains them*. A chain always starts with a point labelled $\ell_2 = E$. By (2a) above, the application of the inertia labelling step of the espalier construction will always use case (i) when γ is labelled E; the chain will *continue* (at least for that step).

(5) *Each chain stops the first time a point* γ' *marked* $\ell_2(\gamma') = P$ *is accepted into the chain*. Consider the competition between cases (i) and (ii) in the step labelled 4, chaining, of the espalier construction. The same type of argument for (2c) with the same value of δ_0 shows that, on the set $s_1 < s < s_1 + \delta_0$, the slope $\dot{v}(s)/\dot{h}(s) > 1$. In running through the list of successors to a point labelled $\ell_2 = P$, one will encounter a point labelled "E" and representing a horizontal intersection before one could potentially encounter a point representing a vertical intersection (which would be needed to continue the chain).

8.2.3. *Proof of edgel approximation lemma*. We consider each property in turn.

Lemma 8.5(i), *all the edgels belonging to* β_m *have endpoints in* $\mathscr{L}(m) \cap \partial B$. This is so by construction, since the endpoints come from the set Γ_2, and $\Gamma_2 \subset \Gamma_1 \subset \mathscr{L}(m) \cap \partial B$.

Lemma 8.5(ii), *each edgel belonging to* β_m *has length less than or equal to* K_1/m. Lemma 8.9(2a) shows that one component of the edgel undergoes displacement exactly $1/m$, while part (2b) shows that the other undergoes displacement no larger than A/m. Hence

$$\text{Length}(e) \le \left(\Delta_1^2 + \Delta_2^2\right)^{1/2} \le \sqrt{A^2 + 1}\,/m < (A + 2)/m = K_1/m.$$

Lemma 8.5(iii), *at least* 50% *of the edgels belonging to* β_m *have length greater than or equal to* $1/m$. Each link in a chain has length greater than or equal to $1/m$, by Lemma 8.9(2a). In between each chain and the next, there is a "switch," an edgel which begins on a vertical (respectively, horizontal) line in $\mathscr{L}(m)$ and ends on a horizontal (respectively, vertical). Since the terminal point of the switch is a point labelled "E," it starts a new chain, of length at least one link. Hence there is at most one "switch" in between each pair of chains. Also, no more than half the edgels in β_m are switches.

Lemma 8.5(iv), *any dyadic square* S *of side* $1/m$ *which intersects* $\{\beta_m\}$ *intersects one and only one edgel*. A dyadic square of side $1/m$ cannot intersect two edgels in the same chain, since each edgel is the only one in that chain which intersects a given column, hence a given square. The square cannot intersect two edgels in different chains. In between any two chains marked H there must be a chain marked V. A chain marked V means that there is a vertical column of thickness at least $1/m$ separating the first H chain from the next H chain.

8.2.4. *Proof of edgelet approximation lemma*. An edgel e in a chain is associated with a column that it traverses completely, in the sense of Lemma 8.9(2a). By part (2b), as e traverses that column it intersects at most K_1 dyadic squares of side $1/m$ in that column. Consider just one of those dyadic squares, S say. The segment $e \cap S$ can be approximated within Hausdorff distance δ by an edgelet $\bar{e} \in \mathscr{E}_{n,\delta}$. To see this, recall the set $V(S)$ defined in Section 2. This consists of vertices at spacing δ around the border of S. It contains two vertices within δ-distance of each endpoint of $e \cap S$; they come just before and just after the corresponding endpoint in a clockwise traverse of the boundary of S. Picking one of the pair at each end yields the endpoints of an edgelet $\bar{e}$. Doing this for each square traversed in the column, and imposing continuity in the choice of approximating edgelets, we get a chain of at most K_1 edgelets approximating the edgel e within Hausdorff distance δ.

Continuing this from edgel to edgel, we obtain a continuous sequence $\overline{E}_m$ of edgelets having cardinality less than or equal to $K_1 \cdot \#E_m$.

In the discretization of edgels to edgelets, we need to attend to certain details. We need to arrange that in all cases the edgelet endpoints are on the same edge of the enclosing dyadic square S as the endpoint of the corresponding $e \cap S$. Also, if the edgel goes exactly through a corner of S, then we must arrange that the edgelet also goes through the corner [this is possible, as the corner belongs to $V(S)$].

When these details are attended to, properties (i)–(iii) follow immediately from the construction. For example, by arranging that the edgelets in $\overline{E}_m$ have endpoints with the same edge incidences as the corresponding $e \cap S$ segments they approximate, we arrange that $\overline{E}_m$ intersects exactly the same dyadic squares of side $1/m$ as the edgels in E_m.

8.2.5. *Counting wedgelets.* Our final step requires the following lemma.

LEMMA 8.10. *Let $B \in$ STAR-SET$^{\alpha}(C)$. Let N_j denote the number of dyadic squares of side 2^{-j} which intersect ∂B. Then, with A_1 and A_2 depending on α and C,*

$$\sum_{j=0}^{J} N_j \le A_1 2^J + A_2, \qquad J \ge 0. \tag{8.11}$$

PROOF. The boundary can be decomposed into a sequence of Lipschitz graphs by the espalier construction. For a Lipschitz graph with constant C the argument of Lemma 5.4 shows that the number of boxes traversed at scale 2^{-j} is at most $(C + 2) \cdot 2^j \cdot R$, where R is the range of the "independent variable" of the graph. The sum of the ranges of the "independent variables" is less than the total arclength of the curve β. The Lipschitz constant of the graph from the espalier construction is an absolute constant A; see Lemma 8.6(ii). The total arclength of β is bounded by a constant L depending on C and α only. Hence at one scale we have

$$N_j \le L \cdot (A + 2) \cdot 2^j.$$

Summing across j gives (8.11) with $A_1 = 2 \cdot L \cdot (A + 2)$.

We now let $N_j(\beta_m)$ denote the number of dyadic squares at scale 2^{-j} intersecting β_m, and note that we have

$$N_j\left(\overline{\beta}_m\right) \le N_j(\beta_m) \le N_j(\beta).$$

We turn now to constructing a recursive dyadic partition based on $\overline{E}_m$. Let $\mathscr{S}_m$ be the collection of dyadic squares of scale $1/m$ that intersect a member of $\overline{E}_m$. Construct the coarsest RDP having the associated members of $\mathscr{S}_m$ as terminal nodes, and decorate the squares belonging to $\mathscr{S}_m$ with the corresponding edgelets, thereby forming an ED-RDP $\overline{P}_m$. We note that we have shown that there really is an ED-RDP with these terminal nodes and decorations, as we have constructed things so that the edgelets are associated with disjoint squares of side $1/m$. Hence Lemma 5.1 applies.

The cardinality of this partition is equal to the number of ancestors and siblings of ancestors of squares of $\mathscr{S}_m$. Now note that any ancestor of a member of $\mathscr{S}_m$ intersects the curve $\overline{\beta}_m$. In the other direction, squares at scale 2^{-j}, $j < \log_2(m)$, whch intersect with this curve will all be ancestors of squares in the ED-RDP.

There are at most three siblings per actual ancestor intersecting the square, so

$$\#\overline{P}_m \le 4 \cdot \sum_{j=0}^{\log_2(m)} N_j\big(\overline{\beta}_m\big) + 1.$$

Applying now (8.11) gives the desired cardinality estimate (8.7). □

9. Discussion.

9.1. *Remarks*.

9.1.1. *Subpixel resolution*. It is important to note that our results specify that we use $\delta = o(1/n)$ to get nearly minimax rates; this corresponds to using edgelets with angular resolution which is asymptotically much finer than what one naively obtains with pixel-level data; there are pairs of edgelets in our dictionary which differ only in their subpixel behavior. The accomodation of subpixel resolution increases the computational complexity of the algorithm, but appears to be necessary.

9.1.2. *Improve log factors*. The log factor given in Theorem 7.1 is not best possible. Using the same arguments as in [15], one sees immediately that the $O(\log(n))$ factor can actually be replaced by $O(\log(n)^r)$, where $r = 2\alpha/(\alpha + 1)$. The same comment applies, for the same reasons, to Theorems 8.1 and 8.2.

9.1.3. *Finer penalization*. In fact the log-factors specified in the results appear to be removable, by changing the wedgelet estimator. Suppose we modify the penalization of partition complexity, so that instead of penalizing cardinality of a partition only, we take notice also of the size of associated dyadic blocks in a partition, and penalize the presence of very small blocks more heavily than large blocks. With an appropriate strategy we can expect to recoup the associated log factors. Indeed, this is a natural analog of the "level-dependent thresholding" idea that recoups log terms in wavelet thresholding (see [20, 22, 3]). Note that we can modify the penalization in this way while still using a fast tree-pruning algorithm.

9.2. *Relations to other work*.

9.2.1. *Breiman–Friedman–Olshen–Stone*. The bulk of the CART book [4] deals with the notion of approximation by recursive partitioning in which "splits" can be made only parallel to the axes. However, in a few places, it also considers more general partitions where "hyperplane splits" are allowed. The ED-RDP partitions discussed in this paper are instances of such general CART partitions; however, they obey very special constraints. We restrict attention to dyadic partitions only ("= midpoint splits" in CARTesian), and we allow only a restricted set of "affine splits" taken from the edgelet dictionary, and we allow such affine splits only on terminal nodes. The value of such restrictions is that they lead to fast algorithms for finding optimal

partitions within this class; and this class of partitions is already large enough to get near-optimal approximations. General recursive partitioning with hyperplane splits seems much too general a model to lead to effective computational algorithms for finding optimal partitions.

9.2.2. *Jones–David–Semmes–Coifman*. The edgelet system we have described here, and the associated dyadic organization of edge data, is closely related to important recent work in harmonic analysis [26, 10]. Peter Jones started off this line of research by showing that one could gather information about approximations to the pieces of a curve defined by intersections with dyadic boxes—recording the error of approximation of such pieces by line segments—could be used to characterize curves of finite arclength (travelling-salesman problem). David and Semmes have extended such dyadic-organization ideas to R^d (where one dissects a surface into dyadically organized pieces and studies approximation by k-planes); they used such tools to understand a number of important questions in analysis.

In this paper, we have focused on a very special discrete set of edgels and showed how to use them to give curves with near-minimax description length. The goal of minimax description length is somewhat different than minimal arclength, although there are significant connections. For our noise-removal purposes it seems to be the description length, rather than arclength, which matters.

Returning to the theme of the Introduction, there is no doubt that the harmonic analysis techniques of David and Semmes will prove to have a variety of applictions to analysis of embedded submanifolds of an ambient Euclidean space. R. R. Coifman is presently working to fashion workable tools for computational harmonic analysis of empirical data. A stimulating meeting organized by Coifman at Yale in December 1996 suggested many potential application areas.

9.3. *Comparison with image processing*. The viewpoint in this paper is, by and large, quite different than the viewpoint one commonly encounters in the field of image processing. In this section we remark on differences and connections.

9.3.1. *Comparison with edge detection*. Existing edge detection methods in the literature of image processing are monoscale—they are typically based on some pixel-scale filtering, followed by contour following. The methods suggested here are explicitly multiscale, and so can involve the use of "detectors" which are of very large scale along an edge. This aspect of our approach causes profound differences, which can be explained as follows. Existing edge detectors can be expected to "work well" provided the noise level is so low that the "edge is visible" at *the pixel scale*. The approach developed here will "work" provided the "edge is obvious" *at some scale*. Here obvious means: the edge is sufficiently straight over a sufficiently large extent

that the signal-to-noise ratio of an edgelet-based line segment detector becomes appreciably bigger than 1.

9.3.2. *About minimax estimation.* The philosophy of minimax estimation is rather foreign to most workers in the image processing literature, and one can reasonably question whether there is a useful contribution of minimaxity to workers in that literature.

Here is an example of what we mean: as the reader no doubt sees clearly, the wedgelets described in this paper are discontinuous. The reconstructions they give will create bad visual artifacts away from the actual boundary being estimated—so-called blocking effects. Such artifacts would be unacceptable in the image processing context. A worker in image processing might argue that, if such methods can be near-optimal, then the optimality is itself suspect.

In our view, this indicates that minimaxity with respect to L^2-loss measures only is a partial goal. A more complete goal would require estimators to satisfy side conditions, or to be minimax with respect to a wider range of losses, which include losses imposing certain visual quality requirements.

We are seeking at the moment a better solution which would use smooth basis elements, obtain near-optimal performance in L^2 and in other losses as well. We expect such a better solution to be based on the ridgelets system deployed in a multiresolution system based on dyadic squares [5]. Ridgelets involve highly elongated basis elements and when deployed in a dyadic scheme, they exhibit certain similarities ot the edgelets scheme.

9.3.3. *About optimal representation.* In the Introduction we claimed that the problem of minimax estimation and the problem of optimal representation are closely linked. From that point of view, the results of this paper show that a certain kind of overcomplete wedgelet system plays the same role for images with edges that sinusoids and wavelets play for the classes of objects traditionally studied in the nonparametric smoothing literature.

There is a branch of the "image analysis" literature which is concerned with issues of optimal representation. Researchers in computational neuroscience have been trying to determine "what are the sparse components of images?" In this body of literature, the aim is to perform an analysis of image data with the goal of uncovering a basis in which typical images have a sparse representation [36]. This is analogous to principal components analysis, only the goal is sparse coefficients rather than uncorrelated coefficients. To the author's eye, the computational results which have been uncovered in this data analysis can be organized in a way reminiscent of the edgelets system and are most easily understood from the edgelets–wedgelets point of view.

The companion article "Sparse Components of Images and Optimal Atomic Decompositions" refers to this interesting work in computational neuroscience and interprets the computational results as being similar to optimal decompositions by wedgelets. The empirically optimal representations of images are very similar to wedgelet decompositions, except that the wedgelets

are discontinuous while the empirically optimal decompositions seem smooth. This seems again to indicate that a next goal would be to remove the discontinuity artifacts from the wedgelets system.

Acknowledgments. This paper was presented at "Asymptotic Methods in Stochastic Dynamics and Nonparametric Statistics," Humboldt University, Berlin, September 2–4, 1996. Thanks to Michael Nussbaum and other organizers of this conference for the invitation to participate.

REFERENCES

[1] BARRON, A. and COVER, T. (1991). Minimum complexity density estimation. *IEEE Trans. Inform Theory* **37** 1034–1054.

[2] BENNETT, N. (1997). Fast algorithms for best anisotropic Walsh bases, and relatives. Ph.D. dissertation, Yale Univ.

[3] BIRGÉ, L. and MASSART, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam* (D. Pollard, E. Torgersen and G. Yang, eds.) 55–89. Springer, New York.

[4] BREIMAN, L., FRIEDMAN, J., OLSHEN, R. and STONE, C. J. (1983). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

[5] CANDÈS, E. and DONOHO, D. (1999). Ridgelets: the key to high-dimensional intermittency? *Philos. Trans. Roy. Soc.* To appear.

[6] CHEN, S., DONOHO, D. L. and SAUNDERS, M. A. (1999). Atomic decomposition by basis pursuit. *SIAM J. Sci Comput.* **20** 33–61.

[7] COIFMAN, R. R., MEYER, Y., QUAKE, S. and WICKERHAUSER, M. V. (1994). Wavelet analysis and signal processing. In *Wavelets and Their Applications* (J. S. Byrnes, J. L. Byrnes, K. A. Hargreaves and K. Berry, eds.) 363–380. Kluwer, Boston.

[8] COIFMAN, R. R. and WICKERHAUSER, M. V. (1992). Entropy-based algorithms for best-basis selection. *IEEE Trans. Inform Theory* **38** 713–718.

[9] DAUBECHIES, I. (1992). *Ten Lectures on Wavelets*. SIAM, Philadelphia.

[10] DAVID, G. and SEMMES, S. (1993). *Analysis of and on Uniformly Rectifiable Sets*. Amer Math Soc., Providence, RI.

[11] DEVORE, R. A. and LORENTZ, G. G. (1993). *Constructive Approximation*. Springer, New York.

[12] DONOHO, D. L. (1993). Unconditional bases are optimal bases for data compression and for statistical estimation. *Appl. Comput. Harmon. Anal.* **1** 100–115.

[13] DONOHO, D. L. (1995). Abstract statistical estimation and modern harmonic analysis. In *Proc. 1994 Internat. Congr. Math.* 997–1005. Birkhäuser, Basel.

[14] DONOHO, D. L. (1996). Unconditional bases and bit-level compression. *Appl. Comput. Harmon. Anal.* **3** 388–392.

[15] DONOHO, D. L. (1997). CART and best-ortho-basis: a connection: *Ann. Statist.* **25** 1870–1911.

[16] DONOHO, D. L. (1998). Sparse components analysis and optimal atomic decomposition. Technical report, Dept. Statistics, Stanford Univ. (http://www-stat.stanford.edu/~donoho/Reports/1998/SCA.ps.)

[17] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **81** 425–455.

[18] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal de-noising in a basis chosen from a library of orthonormal base. *Comp. R. Acad. Sci. Paris A* **319** 1317–1322.

[19] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Empirical atomic decomposition. Unpublished manuscript.

[20] DONOHO, D. L. and JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26** 879–921.

[21] DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B* **57** 301–369.

[22] DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24** 508–539.
[23] FOSTER, D. and GEORGE, E. I. (1994). The risk inflation factor in multiple linear regression. *Ann. Statist.* **22** 1947–1975.
[24] FRAZIER, M., JAWERTH, B. and WEISS, G. (1991). *Littlewood–Paley Theory and the Study of Function Spaces*. Amer. Math. Soc., Providence, RI.
[25] GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation. Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Patt. Anal. Mach. Intel.* **6** 721–741.
[26] JONES, P. W. (1990). Rectifiable sets and the travelling salesman problem. *Invent. Math.* **102** 1–15.
[27] KHAS'MINSKII, R. Z. and LEBEDEV, V. S. (1990). On the properties of parametric estimators for areas of a discontinuous image. *Problems Control Inform. Theory* **19** 375–385.
[28] KOROSTELEV, A. P. (1987). Minimax estimation of a discontinuous signal. *Theory Probab. Appl.* **32** 796–799.
[29] KOROSTELEV, A. P. and TSYBAKOV, A. (1993). *Minimax Theory of Image Reconstruction. Lecture Notes in Statist.* **82**. Springer, New York.
[30] MALLAT, S. (1997). *A Wavelet Tour of Signal Processing*. Academic Press, New York.
[31] MALLAT, S. and ZHANG, Z. (1993). Matching pursuit in a time–frequency dictionary. *IEEE Trans. Signal Processing* **41** 3397–3415.
[32] MARR, D. (1982). *Vision*. Freeman, New York.
[33] MEYER, Y. (1990). *Ondelettes et Operateurs I*: *Ondelettes*. Hermann, Paris. (English translation: *Wavelets and Operators*. Cambridge Univ. Press.)
[34] MEYER, Y. (1993). *Wavelets*: *Algorithms and Applications*. SIAM, Philadelphia.
[35] MÜLLER, H. G. and SONG, K. S. (1994). Maximin estimation of multidimensional boundaries. *J. Multivariate Anal.* **50** 265–281.
[36] OLSHAUSEN, B. A. and FIELD, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381** 607–609.
[37] THIELE, C. M. and VILLEMOES, L. F. (1996). A fast algorithm for adapted time–frequency tilings. *Appl. Comput. Harmon. Anal.* **3** 91–100.
[38] WICKERHAUSER, M. V. (1994). *Adapted Wavelet Analysis from Theory to Software*. Peters, Boston.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
SEQUOIA HALL
STANFORD, CALIFORNIA 94305
E-MAIL: donoho@stat.stanford.edu