The Annals of Statistics
2002, Vol. 30, No. 3, 784–842

RECOVERING EDGES IN ILL-POSED INVERSE PROBLEMS: OPTIMALITY OF CURVELET FRAMES[1]

BY EMMANUEL J. CANDÈS AND DAVID L. DONOHO

Stanford University

We consider a model problem of recovering a function $f(x_1, x_2)$ from noisy Radon data. The function f to be recovered is assumed smooth apart from a discontinuity along a C^2 curve, that is, an edge. We use the continuum white-noise model, with noise level ε.

Traditional linear methods for solving such inverse problems behave poorly in the presence of edges. Qualitatively, the reconstructions are blurred near the edges; quantitatively, they give in our model mean squared errors (MSEs) that tend to zero with noise level ε only as $O(\varepsilon^{1/2})$ as $\varepsilon \to 0$. A recent innovation—nonlinear shrinkage in the wavelet domain—visually improves edge sharpness and improves MSE convergence to $O(\varepsilon^{2/3})$. However, as we show here, this rate is not optimal.

In fact, essentially optimal performance is obtained by deploying the recently-introduced tight frames of *curvelets* in this setting. Curvelets are smooth, highly anisotropic elements ideally suited for detecting and synthesizing curved edges. To deploy them in the Radon setting, we construct a curvelet-based biorthogonal decomposition of the Radon operator and build "curvelet shrinkage" estimators based on thresholding of the noisy curvelet coefficients. In effect, the estimator detects edges at certain locations and orientations in the Radon domain and automatically synthesizes edges at corresponding locations and directions in the original domain.

We prove that the curvelet shrinkage can be tuned so that the estimator will attain, within logarithmic factors, the MSE $O(\varepsilon^{4/5})$ as noise level $\varepsilon \to 0$. This rate of convergence holds uniformly over a class of functions which are C^2 except for discontinuities along C^2 curves, and (except for log terms) is the minimax rate for that class.

Our approach is an instance of a general strategy which should apply in other inverse problems; we sketch a deconvolution example.

1. Introduction. Suppose we wish to recover an object $f(t)$—a function in $L^2(\mathbf{R}^d)$—but we are able to observe data only about $g(u) = (Kf)(u)$, where K is a linear transformation, such as a Radon transform or convolution transform. Such

Received June 2000; revised November 2001.

[1] Supported in part by NSF Grants DMS-98-72890 (KDI) and DMS-95-05151 and by AFOSR MURI-95-P49620-96-1-0028.

AMS 2000 subject classifications. Primary 62G20, 62C20; secondary 41A30, 42C40, 94A08, 94A12.

Key words and phrases. Ill-posed inverse problems, regularization, singular value decomposition, Radon transform, deconvolution, edge, edge-preserving regularization, wavelets, curvelets, ridgelets, wavelet-vaguelette decomposition, wavelet shrinkage, minimax estimation, optimal rates of convergence, white-noise model.

linear inverse problems arise in scientific settings ranging from medical imaging to physical chemistry to extragalactic astronomy. Moreover, we assume that the data are noisy, so that we observe $y(u)$ given by

$$y(u) = (Kf)(u) + z(u), \qquad u \in \mathcal{U},$$

where z is a noise (whether stochastic or deterministic, we do not as yet specify). We are interested in recovering f from the data y. For definiteness, we use the $L^2(\mathbf{R}^d)$ norm $\|\hat{f} - f\|_2$ to measure quality of recovery.

One's first impulse might be to attempt the estimate $\hat{f} = K^{-1}y$. However, in the cases of most interest scientifically, K is not invertible, in the sense that K^{-1} does not exist as a bounded linear operator; such inverse problems are called ill-posed. For reviews of these concepts, see Bertero [1], O'Sullivan [55] and Wahba [64].

It is now standard to approach inverse problems by the method of regularization [62], in which one applies, rather than K^{-1}, a linear operator of the form $(K^T K + \lambda\Sigma)^{-1} K^T$. This typically produces a reconstruction in which certain features of the original are "smoothed away." This phenomenon is very evident in imaging applications, such as medical and seismic, where often the reconstructions by the method of regularization are seen to be blurred versions of the original.

This blurring phenomenon is particularly of concern when the underlying object has edges and when the location and size of these edges are of central interest. Such edge-dominated situations are relatively common in imaging applications, where the edges signify boundaries between different parts of a scene or different layers in the earth, or different organs in the body. It would be of interest to obtain sharper reconstructions for objects with edges.

The phenomenon of blurring and the goal of edge recovery have been studied by many researchers over the last few years; a partial listing of articles specifically devoted to this theme would include [10, 14, 38, 39, 45, 58, 60, 61].

Many creative ideas have been brought to bear on this problem, including Markov random fields, anisotropic diffusions, level-set methods, total variation regularization methods and nonconvex optimization techniques.

Many of the cited articles propose heuristically valuable methods, which, when applied to concrete imaging problems at currently practical scales of resolution and noise, produce visually appealing results using currently available computing resources. However, in our opinion, much of the cited work, though practically valuable, lacks a theoretical perspective that would allow one to say that the problem is really well understood. What are the ultimate limits of performance in recovering objects with edges from indirect, noisy observations? What methods can attain that performance?

In this article, we develop a theoretical perspective on these questions using tools from harmonic analysis and statistical decision theory. We are able to decompose the inverse problem in a new way and obtain insights about the degree of attainable performance which are quite different from those which might be inferred from the literature cited above.

In this Introduction we develop the theme of decomposing an inverse problem, both from a classical viewpoint and from a new viewpoint and discuss insights derivable from these decompositions, leading up to a statement of our main results. We hope that many readers will be able to follow us through the introduction, even if later sections involve analysis of a type they do not intend to pursue in detail.

1.1. *The SVD paradigm.* It is now standard to decompose inverse problems by *singular value decomposition* methods, defined as follows. We let $\|\cdot\|_2$ stand equally for the $L^2(dt)$ and $L^2(du)$ norms, and let $\langle , \rangle$ and $[,]$ denote the respective inner products. If K^*K is a compact operator we let $(e_\nu(t))$ denote its eigenfunctions, k_ν^2 its eigenvalues, and $h_\nu(u)$ the normalized image $h_\nu(u) = (Ke_\nu)(u)/\|Ke_\nu\|$ of these. If no k_ν is zero, we have the reproducing formula

$$f = \sum_\nu k_\nu^{-1}[Kf, h_\nu]e_\nu.$$

A reconstruction rule may be based on this formula, and the idea that the "important" coefficients $\langle f, e_\nu\rangle$ occur early in the series. Then, picking weights w_ν which are near 1 for small ν and near 0 for large ν we get a *Windowed SVD* reconstruction formula:

$$\hat{f}_w = \sum_\nu w_\nu k_\nu^{-1}[y, h_\nu]e_\nu. \tag{1.1}$$

Weights are chosen so that $(w_\nu/k_\nu) \in \ell^2$. As the eigenvalues of the compact operator K^*K tend to zero this weighting is necessary so that division by near-zero elements does not prevent convergence of the series.

The windowed SVD method, at least theoretically, includes many other approaches to inversion as special cases, simply by suitable choice of the window function w_ν; see [1] for example. Thus, if we pick $w_\nu = \frac{k_\nu^2}{k_\nu^2+\lambda}$, we get the method of regularization, and if we pick $w_\nu = (1 - (1 - \mu k_\nu^2)^m)$ we get the mth iterative damped backprojection [1].

The singular system decomposition has led to applications in a variety of fields. See Bertero, De Mol, and Pike [2] for a physics-oriented treatment, and Johnstone and Silverman [42, 43] for a statistics-oriented example. In fact, the intensive work by many researchers building an extensive edifice of SVD applications qualifies the SVD as a *paradigm* for analyzing and solving linear inverse problems.

1.2. *Limitations of SVD.* Despite the great popularity of schemes based explicitly or implicitly on SVD, the method suffers from performance limitations. These are rooted in the fact that the basis functions (e_ν), (h_ν) derive from the operator under study, not from the object to be recovered. Thus, if the same operator K occurs in two different fields of scientific inquiry, the basis functions will be the same, even though the type of object to be recovered may be quite different in the two fields. One can easily imagine that in one field of scientific inquiry the f to be recovered could be very efficiently represented in terms of the basis set used, while in the other area, the object is poorly approximated by finite sums of basis terms e_ν even when a fairly large number of terms is used.

Efficient representation of the object f by singular functions e_ν is essential. Suppose that (for definiteness) the object is observed in white noise, so that the observed singular coefficient obeys

$$[y, h_\nu] = k_\nu \theta_\nu + \varepsilon z_\nu,$$

where (z_ν) is a Gaussian white noise sequence, ε is the noise level, and $\theta_\nu = \langle e_\nu, f \rangle$ is the component of f in the direction e_ν. Then, if we use the best window (w_ν) possible for the function under consideration we would have a mean-squared error within a factor of 2 of

$$\sum_\nu \min(\theta_\nu^2, k_\nu^{-2}\varepsilon^2), \tag{1.2}$$

which we take as a proxy for the difficulty in recovering f. This expression shows that in order to have accurate reconstructions, it is important that there be very few θ_ν which are large, and that those which are large be located at those components ν where k_ν is also large.

In short, even when the SVD window (w_ν) is chosen optimally for the specific function at hand, it is necessary for the coefficients (θ_ν) to have a certain distribution of energy in the singular system basis. Otherwise, the windowed SVD method will have poor MSE properties.

In many realistic examples one does *not* have the desired agreement between the energy distribution of the object and the decay of the singular values. Suppose we are considering a two-dimensional inverse problem involving deconvolution, and suppose we impose circular boundary condition so that the singular functions are known explicitly—they are sinusoids. Suppose that the object to be recovered has a discontinuity along a smooth curve. Then its Fourier coefficients θ_ν typically decay as $|\nu| \to \infty$ only like $1/|\nu|^{3/2}$, which is rather slow; in consequence, the expression $\sum_\nu \min(\theta_\nu^2, k_\nu^{-2}\varepsilon^2)$ will tend to zero slowly with ε.

This is a general phenomenon and continues outside the special case of deconvolution. Whenever the object to be recovered has edges, and the SVD has sinusoidal structure, SVD-based approaches, and cognate approaches such as damped backprojection and the method of regularization, will have trouble.

In many typical cases, SVD-based methods are nearly optimal among linear methods, so we can infer that if the SVD-based methods have poor MSE properties, so will other linear methods.

1.3. *Ubiquity of edges.* Objects with discontinuities along edges arise in many important inverse problems arising in imaging applications.

For example, in seismic inverse problems, the object to be recovered represents bulk material properties as a function of depth, and so can be expected to change discontinuously across layer boundaries. Geophysicists have used "layer cake" models of the earth with considerable success for many years.

In biomedical imaging, the object to be recovered might represent either the material density or the metabolic activity as a function of space; discontinuities represent changes in material–metabolic properties across organ boundaries. Biomedical imaging researchers have used piecewise constant "phantoms" in evaluating their imaging algorithms for years, with good success. They seem to regard piecewise constant imagery as a reasonable starting model even today.

Notice that the typical model in such applications concentrates essentially all the information in the edges, and yet such edge-dominated objects are precisely the type causing problems for the standard SVD methods for reconstructing inverse problems. This mismatch, as mentioned above, is at the source of a considerable body of recent research.

1.4. *An alternative strategy.* The SVD focuses exclusively on properties of the forward operator rather than the object to be recovered. In essence, it diagonalizes the forward operator. There is an interesting alternative strategy. Using language from harmonic analysis, which we will now begin to employ more and more heavily, the strategy is *to develop a new decomposition which is much better adapted to the type of edge-dominated* object *we need to recover while providing an almost diagonal representation of the* operator. Below, we argue that such a strategy may lead to algorithms which, *for typical edge-dominated objects of interest, enjoy dramatically lower MSE than the classical SVD approach*.

In this article, we consider as a model problem a mathematical caricature of image processing, the case where typical objects are functions of two variables with discontinuities along edges and which are otherwise smooth. We develop decompositions of both the inverse problem and the object to be recovered using a newly developed tool: frames of curvelets. These new frames serve almost as well as the SVD for diagonalizing the forward operator K of certain inverse problems, while doing much better for representing objects with edges.

We consider specifically the problem of noisy Radon inversion where the data are obtained in the so-called white noise model,

$$Y(dt, d\theta) = (Rf)(t, \theta)\, dt\, d\theta + \varepsilon W(dt, d\theta). \tag{1.3}$$

Here R denotes the Radon transform, taking functions $f(x_1, x_2)$ on $\mathbf{R}^2$ into functions $(Rf)(t, \theta)$ on $\mathbf{R} \times [-\pi, \pi)$ formed by integration along lines of codirection θ and distance t from the origin; $W(\theta, t)$ denotes a Wiener sheet (i.e., the primitive of white noise); ε is a noise level and f is the object to be recovered. For SVDs of the Radon transform, see [12, 35]. Over the last decade, the white-noise model has proven to be a fruitful theoretical tool. Although the model is continuous and real data are typically discretely sampled, the asymptotic theory deriving from the white-noise model has typically been found to lead directly to comparable asymptotic theory in a sampled data model. Without belaboring this point we can cite general theory [3, 54], examples of the very clean derivations of optimal procedures possible in the white-noise model [16, 18, 20, 24, 27, 30, 41, 56, 57] and many successful applications of this principle to sampled data [17, 28, 29, 31–33, 54]. (We are of course aware of [34].)

For some readers this model may seem initially rather remote; they may be helped by the observation that what it really says is: each integral $\int \nu(t,\theta)Y(dt\,d\theta)$ of the observed data Y is normally distributed with mean $\int \nu(t,\theta)f(t,\theta)\,dt\,d\theta$ and variance $\varepsilon^2 \int \nu(t,\theta)^2\,dt\,d\theta$.

We develop an operator-biorthogonal decomposition of the Radon transform based on curvelet frames. Using this decomposition, we propose a method that yields dramatic benefits in asymptotic mean-squared error over previous approaches, and in fact a near-optimality. In the remainder of the introduction we develop further background for stating our result.

1.5. *Objects with singularities along curves.* Suppose we have an object f supported in $[0,1]^2$ which is smooth away from a discontinuity across a C^2 curve. It is well known that, in this setting, wavelets offer an improvement on traditional representations like sinusoids, but wavelets are far from optimal.

To make this concrete, consider approximating such an f from the best m-terms in a Fourier expansion adapted to $[-\pi,\pi]^2$ (say). The squared error of such an m-term expansion $\tilde{f}_m^F$ would obey

$$\|f - \tilde{f}_m^F\|_2^2 \asymp m^{-1/2}, \qquad m \to \infty. \tag{1.4}$$

For comparison, consider an approximation $\tilde{f}_m^W$ from the best m-terms in a wavelet expansion; then

$$\|f - \tilde{f}_m^W\|_2^2 \asymp m^{-1}, \qquad m \to \infty, \tag{1.5}$$

which is considerably better. However, from [21, 23, 26] we know that there exist dictionaries of (nonorthogonal) elements, and procedures for selecting from those dictionaries that will yield m-term approximations obeying

$$\|f - \tilde{f}_m^D\|_2^2 \asymp m^{-2}, \qquad m \to \infty. \tag{1.6}$$

The relative disadvantage of Fourier and wavelets methods for purposes of efficient representation has an immediate counterpart in solving inverse problems. Suppose we have data from the Radon transform with white noise, (1.3), again with a smooth f having a discontinuity along a generic C^2 curve. If we apply the SVD in this setting, we get the result

$$MSE(\text{WINDOWED SVD}, f) \asymp \varepsilon^{1/2}, \qquad \varepsilon \to 0.$$

In recent work described further below, [18] proposed a wavelet-based method for solving inverse problems of this kind, called the *wavelet-vaguelette decomposition*. If we apply the WVD in this setting (see also Section 9 below) we get an improvement over SVD,

$$MSE(\text{Thresholded WVD}, f) \asymp \varepsilon^{2/3}, \qquad \varepsilon \to 0.$$

However, results farther below in this paper establish that the minimax mean-squared error will in this setting be $O(\varepsilon^{4/5-\delta})$ for each $\delta > 0$. Hence, when we consider inverse problems involving otherwise smooth objects having discontinuities along edges, a representation based on wavelets, though an improvement on linear or SVD methods, is substantially suboptimal.

1.6. *Sparse representations by curvelets.* In recent work [8, 9], we introduced tight frames of curvelets, systems with the following properties:

1. There is a collection (γ_μ) with μ running through a discrete index set $\mathcal{M}$ which makes a tight frame for $L^2(\mathbf{R}^2)$. This means there is a reproducing formula
$$f = \sum_\mu \langle f, \gamma_\mu\rangle \gamma_\mu$$
and a Parseval-type relation,
$$\|f\|^2_{L^2(\mathbf{R}^2)} = \sum_\mu |\langle f, \gamma_\mu\rangle|^2.$$
2. The set $\mathcal{M}$ has a seven-index structure $\mu = (s, k_1, k_2; j, k; i, \ell, \varepsilon)$ to be described below, whose indices include parameters for *scale*, *location*, *direction* and *microlocation*.
3. The elements of the tight frame with substantial L^2-norm obey a special "anisotropic scaling law": the width of the effective support of these elements is effectively proportional to the length *squared*. The frame elements become successively more anisotropic at progressively finer scales.
4. The number of distinct directions at a given scale grows as *scale*$^{-1}$.

The last two properties are quite different from those of preexisting multiscale representations such as wavelets, where the aspect ratios of basis–frame elements remain fixed as one moves to finer scales and the number of distinct directions remains fixed also.

The construction is briefly reviewed in Section 2 below. Details of the construction and heuristic insights are provided in [8, 9].

Our motivation leading to this construction (and to the choice of name) was the problem of representing otherwise smooth objects which have a discontinuity along a C^2 curve. In [8] it is shown that if f has C^2 smoothness away from a simple discontinuity along a C^2 curve, the curvelet coefficients $\alpha_\mu = \langle f, \gamma_\mu\rangle$ obey

$$\sum_\mu |\alpha_\mu|^p < \infty$$

for each $p > 2/3$. It follows from this and the tight frame property that for such f, m-term curvelet approximations $\tilde{f}_m^C$ made using the m terms with biggest curvelet coefficients obey, for each $\delta > 0$,

$$\|f - \tilde{f}_m^C\|_2 \leq C_\delta m^{-2+\delta}, \qquad m \to \infty.$$

Comparing with the results (1.4)–(1.6) we see that m-term approximations in the curvelet frame are almost rate optimal, and in fact perform far better than m-term sinusoid or wavelet approximations, in an asymptotic sense.

1.7. *Operator-biorthogonal curvelet decomposition.* Obviously, the curvelet representation is far more effective in representing objects with edges than wavelets or more traditional representations. Interestingly, curvelets also afford an almost-diagonal representation of the Radon operator, or equivalently its Gram operator R^*R.

The operator-biorthogonal curvelet decomposition (BCD) has the following ingredients, stated for a general operator K, of which the Radon operator R is a particular case.

1. We start with (γ_μ) a curvelet tight frame for $L^2(dx_1\,dx_2)$. This is intended to play the part that was played in the SVD theory by the eigenfunctions (e_ν). In short, curvelets are used in place of singular functions.
2. Based on the forward operator K and our choice of (γ_μ), we obtain systems (U_μ) and (V_μ) in $L^2(dt\,d\theta)$.
3. In fact the systems are generated according to the relations

   $$K\gamma_\mu = \kappa_s V_\mu, \qquad K^* U_\mu = \kappa_s \gamma_\mu.$$

   Here, the scalars κ_s are defined by certain scaling properties of the operator K, and are called quasi-singular values; in effect they normalize the functions $K\gamma_\mu$ and $(K^*)^{-1}\gamma_\mu$, so that the systems (U_μ) and (V_μ) obey $\|U_\mu\| \asymp 1$, $\|V_\mu\| \asymp 1$.
4. The systems (U_μ) and (V_μ) obey the generalized biorthogonality relations

   $$[U_\mu, V_{\mu'}] = 2^{s'-s}\langle \gamma_\mu, \gamma_{\mu'} \rangle.$$

5. The systems are frames: for (U_μ) we have

   $$\left(\sum_\mu \langle U_\mu, g\rangle^2\right)^{1/2} \asymp \|g\|_{L^2(dt\,d\theta)},$$

   $$\left\|\sum_\mu a_\mu U_\mu\right\|_2 \leq C\|(a_\mu)\|_{\ell^2}$$

   and similarly for (V_μ).

To summarize: the BCD may be viewed as an analog of the SVD in which curvelets play the role of the eigenbasis e_ν and dual curvelets U_μ and V_μ play the role of the dual functions h_ν and the κ_s play the role of approximate singular values.

An immediate consequence of the BCD construction is the *reproducing formula*,

$$f = \sum_\mu [Rf, U_\mu]\kappa_s^{-1}\gamma_\mu, \tag{1.7}$$

which makes sense for every f which is a finite linear combination of γ_μ's. This shows that the curvelet coefficient $\langle \gamma_\mu, f\rangle$ can be obtained from noiseless Radon data Rf by simply using the U_μ frame coefficient.

This formula directly associates behavior in the Radon domain with behavior in the object domain. The Radon domain data are analyzed by a bank of functions (U_μ) and each output coefficient scales a corresponding synthesized behavior γ_μ in the object domain. Since the γ_μ correspond at fine scales to highly localized directionally oriented elements, the formula may be said at fine scales to be reading off the existence of edges at certain locations and orientations in the object domain from behavior of the Radon transform.

1.8. *Statistical estimation.* In the reproducing formula (1.7), the κ_s in (1.7) are tending to zero as $s \to \infty$, so the reproducing formula is very sensitive to the presence of nonzero terms at large values of s. In particular, it would be rather foolish to use this formula *as is* on inaccurate data. For dealing with noisy data, we propose a rule of the general form

$$\hat{f} = \sum_\mu \delta([Y, U_\mu]\kappa_s^{-1}, t_s)\gamma_\mu,$$

where $\delta(\cdot, t)$ is a scalar thresholding nonlinearity with threshold t, and t_s are appropriate scale-dependent thresholds.

This makes sense; because the curvelet transform has its big coefficients at unpredictable locations (depending on the location of the edge curve), we cannot say a priori where the “important coefficients” will be; therefore we apply thresholding.

We are able to obtain the following result for Radon inversion in the white noise model (1.3). *Suppose that f is compactly supported and C^2 smooth away from a C^2 curve. Then for each $\delta > 0$,*

$$MSE(\text{Thresholded BCD}, f) = O(\varepsilon^{4/5-\delta}), \qquad \varepsilon \to 0. \tag{1.8}$$

A heuristic explanation for the exponent $4/5$ is given in Section 6 below.

As we have seen, linear SVD damping methods and nonlinear wavelet shrinkage methods achieve MSE convergence rates $O(\varepsilon^{1/2})$ and $O(\varepsilon^{2/3})$, respectively. Thus the curvelet-based approach to Radon inversion can substantially outperform existing methods.

In fact we obtain a stronger result, valid uniformly over a class of such f. The proof is given in Section 7. It follows from a lower bound we develop in Section 8 that no measurable procedure can possibly achieve better than a result of order $\varepsilon^{4/5}$ in this problem. The rate result (1.8) is therefore essentially unimprovable for a class of otherwise smooth objects with edges.

1.9. *Contents.* Section 2 reviews the construction of curvelets. Section 3 constructs a stable biorthogonal decomposition of the Laplacian based on curvelets. Based on this, Section 4 constructs the operator-biorthogonal decomposition of the Radon transform. Section 5 interprets the resulting dual analyzing elements. Section 6 gives a heuristic indicating why the $4/5$ law may be expected to hold. Section 7 gives the proof of our main result, a strengthening of (1.8). Section 8 proves that no result better than this can be expected, uniformly over a class of objects with edges. Section 9 discusses generalizations, for example, to deconvolution problems; we believe that results of the kind proved here hold for a wide variety of inverse problems. It also points out that despite the existence of a rather voluminous literature on "edge-preserving methods," no previously known methods seem to approach the $4/5$ law. Finally, proofs of key estimates supporting our main result are given in Section 10.

2. Curvelet construction.

We now briefly discuss the curvelet frame; for more details, see [8]. The construction combines several ingredients, which we briefly review.

1. *Ridgelets*, a method of analysis very suitable for objects which are discontinuous across straight lines;
2. *Multiscale ridgelets*, a pyramid of analyzing elements which consists of ridgelets renormalized and transported to a wide range of scales and locations;
3. *Bandpass filtering*, a method of separating an object out into a series of disjoint scales.

We briefly describe each component in turn, and then their combination. There is a difference between this construction and the one given in [8] at large scales.

2.1. *Ridgelets.* The theory of ridgelets was developed in the Ph.D. thesis of Emmanuel Candès [5, 4]. In that work, Candès showed that one could develop a system of analysis based on ridge functions

$$\psi_{a,b,\theta}(x_1, x_2) = a^{-1/2}\psi\big((x_1\cos(\theta) + x_2\sin(\theta) - b)/a\big). \tag{2.1}$$

He introduced a continuous ridgelet transform $R_f(a, b, \theta) = \langle\psi_{a,b,\theta}(x), f\rangle$ with a reproducing formula and a Parseval relation. He showed how to construct frames, giving stable series expansions in terms of a special discrete collection of ridge functions. The approach was general, and gave ridgelet frames for functions in $L^2[0, 1]^d$ in all dimensions $d \geq 2$. For further developments, see [5–7].

Reference [22] showed that in two dimensions, by heeding the sampling pattern underlying the ridgelet frame, one could develop an orthonormal set for $L^2(\mathbf{R}^2)$ having the same applications as the original ridgelets. The orthoridgelets are indexed using $\lambda = (j, k, i, \ell, e)$, where j indexes the ridge scale, k the ridge location, i the angular scale, and ℓ the angular location; e is a gender token. Roughly speaking, the orthoridgelets look like pieces of ridgelets (2.1) which are windowed to lie in discs of radius about 2^i; $\theta_{i,\ell} = \ell/2^i$ is roughly the orientation parameter, and 2^{-j} is roughly the thickness.

Throughout the paper $\xi \in \mathbf{R}^2$ will index bidimensional frequencies and in the frequency plane we will often make use of polar notation (ω, θ),

$$\xi(\omega, \theta) = (\omega \cos\theta, \omega \sin\theta), \qquad \omega > 0,\ \theta \in [0, 2\pi).$$

A formula for orthoridgelets can be given in the frequency domain (the notation $\hat{\cdot}$ is used for the Fourier transform)

$$\hat{\rho}_\lambda\big(\xi(\omega,\theta)\big) = \omega^{-1/2}\big(\hat{\psi}_{j,k}(\omega) w^e_{i,\ell}(\theta) + \hat{\psi}_{j,k}(-\omega) w^e_{i,\ell}(\theta + \pi)\big)/2. \tag{2.2}$$

Here the $\psi_{j,k}$ are Meyer wavelets for $\mathbf{R}$ [49, 52], $w^e_{i,\ell}$ are periodic wavelets for $[-\pi, \pi)$, indices run as follows: $j, k \in \mathbf{Z}$, $\ell = 0, \dots, 2^{i-1} - 1$; $i \geq i_0$, and, if $e = 0$, $i = \max(i_0, j)$, while if $e = 1$, $i \geq \max(i_0, j)$; i_0 is an arbitrary nonnegative integer that we may take to be zero. Notice the restrictions on the range of i, ℓ. Let Λ denote the set of all such indices λ. (For general information about wavelets, see [51, 53].)

2.2. *Multiscale ridgelets.* Think of orthoridgelets as objects which have a "length" of about 1 and a "width" which can be arbitrarily fine. The multiscale ridgelet system renormalizes and transports such objects, so that one has a system of elements at all lengths and all finer widths.

The construction begins with a smooth partition of energy function $w(x_1, x_2) \geq 0$, $w \in C_0^\infty([-1, 1]^2)$ obeying $\sum_{k_1,k_2} w^2(x_1 - k_1, x_2 - k_2) \equiv 1$. Define a transport operator, so that with index Q indicating a dyadic square $Q = (s, k_1, k_2)$ of the form $[k_1/2^s, (k_1 + 1)/2^s) \times [k_2/2^s, (k_2 + 1)/2^s)$, by $(T_Q f)(x_1, x_2) = f(2^s x_1 - k_1, 2^s x_2 - k_2)$. The *multiscale ridgelet* with index $\mu = (Q, \lambda)$ is then

$$\psi_\mu = 2^s T_Q(w\rho_\lambda).$$

In short, one transports the normalized, windowed orthoridgelet.

Letting $\mathcal{Q}_s$ denote the dyadic squares of side 2^{-s}, we can define the subcollection of *monoscale ridgelets* at scale s:

$$\mathcal{M}_s = \big\{(Q, \lambda) : Q \in \mathcal{Q}_s, \lambda \in \Lambda\big\}.$$

It is immediate from the orthonormality of the ridgelets that each system of monoscale ridgelets makes tight frame, in particular obeying the Parseval relation

$$\sum_{\mu \in \mathcal{M}_s} \langle \psi_\mu, f \rangle^2 = \|f\|^2_{L^2}.$$

It follows that the dictionary of multiscale ridgelets at all scales, indexed by

$$\mathcal{M} = \bigcup_{s \geq 1} \mathcal{M}_s$$

is not frameable, as we have energy blow-up,

$$\sum_{\mu \in \mathcal{M}} \langle \psi_\mu, f \rangle^2 = \infty. \tag{2.3}$$

The multiscale ridgelets dictionary is simply too massive to form a good analyzing set. It lacks interscale orthogonality; $\psi_{(Q,\lambda)}$ is not typically orthogonal to $\psi_{(Q',\lambda')}$ if Q and Q' are squares at different scales and overlapping locations. In analyzing a function using this dictionary, the repeated interactions with all different scales causes energy blow-up (2.3).

The construction of curvelets solves this problem by in effect disallowing the full richness of the multiscale ridgelets dictionary. Instead of allowing all different combinations of "lengths" and "widths," we allow only those where *width* $\approx$ *length*^{2}.

2.3. *Subband filtering.* Our remedy to the "energy blow-up" (2.3) is to decompose f into subbands using standard filterbank ideas. Then we assign one specific monoscale dictionary $\mathcal{M}_s$ to analyze one specific (and specially chosen) subband.

We define coronae of frequencies $|\xi| \in [2^{2s}, 2^{2s+2}]$, and subband filters D_s extracting components of f in the indicated subbands; a filter P_0 deals with frequencies $|\xi| \leq 1$. The filters decompose the energy exactly into subbands:

$$\|f\|_2^2 = \|P_0 f\|_2^2 + \sum_s \|D_s f\|_2^2.$$

The construction of such operators is standard [63]; the coronization oriented around powers 2^{2s} is nonstandard and essential for us. Explicitly, we build a sequence of filters Φ_0 and $\Psi_{2s} = 2^{4s}\Psi(2^{2s}\cdot)$, $s = 0, 1, 2, \ldots$, with the following properties: Φ_0 is a lowpass filter concentrated near frequencies $|\xi| \leq 1$; Ψ_{2s} is bandpass, concentrated near $|\xi| \in [2^{2s}, 2^{2s+2}]$ and we have

$$|\hat{\Phi}_0(\xi)|^2 + \sum_{s \geq 0} |\hat{\Psi}(2^{-2s}\xi)|^2 = 1 \qquad \forall \xi.$$

Hence, D_s is simply the convolution operator $D_s f = \Psi_{2s} * f$.

2.4. *Definition of curvelet transform.* Assembling the above ingredients, we are able to sketch the definition of the curvelet transform. We let M' consist of M merged with the collection of integral triples (s, k_1, k_2, e) where $s \leq 0$, $e \in \{0, 1\}$, indexing all dyadic squares in the plane of side $2^s > 1$.

The curvelet transform is a map $L^2(\mathbf{R}^2) \mapsto \ell^2(M')$, yielding curvelet coefficients $(\alpha_\mu : \mu \in M')$. These come in two types.

At *coarse scales* we have wavelet coefficients:

$$\alpha_\mu = \langle W_{s,k_1,k_2,e}, P_0 f\rangle, \qquad \mu = (s, k_1, k_2) \in M' \backslash M,$$

where each $W_{s,k_1,k_2,e}$ is a Meyer wavelet, while at *fine scale* we have multiscale ridgelet coefficients of the bandpass filtered object:

$$\alpha_\mu = \langle D_s f, \psi_\mu\rangle, \qquad \mu \in M_s, s = 1, 2, \dots .$$

Note well that for $s > 0$, each coefficient associated to scale 2^{-s} derives from the subband filtered version of f—$D_s f$—and not from f.

Several properties are immediate:

1. Tight frame:

$$\|f\|_2^2 = \sum_{\mu \in M'} |\alpha_\mu|^2.$$

2. Existence of coefficient representers (frame elements): there are $\gamma_\mu \in L^2(\mathbf{R}^2)$ so that

$$\alpha_\mu \equiv \langle f, \gamma_\mu\rangle.$$

3. L^2 reconstruction formula:

$$f = \sum_{\mu \in M'} \langle f, \gamma_\mu\rangle \gamma_\mu.$$

4. Formula for frame elements: for $s \leq 0$, $\gamma_\mu = P_0 \phi_{s,k_1,k_2}$, while for $s > 0$,

$$\gamma_\mu = D_s \psi_\mu, \qquad \mu \in \mathcal{Q}_s. \tag{2.4}$$

In short, fine-scale curvelets are obtained by bandpass filtering of multiscale ridgelet coefficients where the *passband* is rigidly linked to the *scale* of spatial localization.

5. Anisotropy scaling law: by linking the filter passband $|\xi| \approx 2^{2s}$ to the scale of spatial localization 2^{-s} imposes that (i) most curvelets are negligible in norm (most multiscale ridgelets do not survive the bandpass filtering D_s); (ii) the nonnegligible curvelets obey *length* $\approx 2^{-s}$ while *width* $\approx 2^{-2s}$. In short, the system obeys approximately the scaling relationship

$$\textit{width} \approx \textit{length}^2.$$

Note: it is at this last step that our 2^{2s} coronization scheme comes fully into play;

6. Oscillatory nature. Both for $s > 0$ and $s \leq 0$, each frame element has a Fourier transform supported in an annulus away from 0.

3. Powers of the Laplacian. It is well known that the Radon transform is intimately involved with certain homogeneous Fourier multiplier operators often called "fractional powers of the Laplacian." In this section, we study the decomposition of such operators by the curvelet frame.

Now the usual Laplacian $\Delta = \sum_{i=1}^{2} \frac{\delta^2}{\delta x_i^2}$ corresponds to the Fourier multiplier $(\Delta f)\hat{}(\xi) = -|\xi|^2 \hat{f}(\xi)$; it makes sense therefore to define the α-power of the Laplacian by

$$((-\Delta)^{\alpha} f)\hat{}(\xi) = |\xi|^{2\alpha} \hat{f}(\xi).$$

Define now, for a curvelet $\gamma_\mu(x_1, x_2)$, two *companions* $\gamma_\mu^{\pm}(x_1, x_2)$ according to

$$\gamma_\mu^{\pm} = 2^{\mp s}(-\Delta)^{\pm 1/4}\gamma_\mu,$$

where, of course, s refers to the scale index occupying the first slot $(s, k_1, k_2, j, k, i, \ell, e)$ in the curvelet index μ. Because γ_μ is effectively concentrated in the frequency domain near $|\xi| \approx 2^{2s}$, we have $2^{2s}|\xi| \approx 1$ through the bulk of the frequency domain support of γ_μ and hence we anticipate $\|\gamma_\mu^{\pm}\| \approx \|\gamma_\mu\|$.

THEOREM 1. *The systems $(\gamma_\mu^{+})_{\mu\in\mathcal{M}}$ and $(\gamma_\mu^{-})_{\mu\in\mathcal{M}}$ are frames for $L^2(\mathbf{R}^2)$: either (fixed) choice of sign ± gives a system with ℓ^2 stable synthesis*

$$\left\| \sum_\mu a_\mu \gamma_\mu^{\pm} \right\|_2 \le C \left(\sum_\mu a_\mu^2 \right)^{1/2} \qquad \forall\, (a_\mu) \in \ell^2, \tag{3.1}$$

and L^2-norm equivalence

$$\left\| \sum_\mu \langle f, \gamma_\mu^{\pm} \rangle \right\|_2 \asymp \|f\|_2 \qquad \forall\, f \in L^2(\mathbf{R}). \tag{3.2}$$

Moreover, the two systems are quasi-biorthogonal:

$$\langle \gamma_\mu^{+}, \gamma_{\mu'}^{-} \rangle = 2^{s'-s} \langle \gamma_\mu, \gamma_{\mu'} \rangle, \qquad \mu, \mu' \in \mathcal{M}, \tag{3.3}$$

where $\langle \gamma_\mu, \gamma_{\mu'} \rangle$ is the reproducing kernel of the curvelet tight frame.

PROOF. We first consider (3.3). Passing to the frequency domain, we have

$$\int \hat{\gamma}_\mu^{+}(\xi)\overline{\hat{\gamma}_{\mu'}^{-}(\xi)}\, d\xi = \int 2^{-s}|\xi|^{1/2}\hat{\gamma}_\mu(\xi) 2^{s'}|\xi|^{-1/2}\overline{\hat{\gamma}_{\mu'}(\xi)}\, d\xi.$$

Canceling the offsetting $|\xi|$-multipliers gives

$$\langle \gamma_\mu^{+}, \gamma_{\mu'}^{-} \rangle = 2^{s'-s} \langle \gamma_\mu, \gamma_{\mu'} \rangle$$

and we note that $\langle \gamma_\mu, \gamma_{\mu'} \rangle = 0$ unless $|s - s'| \le 1$, completing the proof of (3.3). The rapid decay of the curvelet frame Gram matrix $\langle \gamma_\mu, \gamma_{\mu'} \rangle$ justifies the terminology "quasi-biorthogonal."

We now turn to the norm equivalence (3.2). Let $\hat{w}_s(\xi)$ be a real-valued smooth radial window supported in $\Xi_{s-2} \cup \cdots \cup \Xi_{s+2}$ which is equal to one on $\xi \in \Xi_{s-1} \cup \Xi_s \cup \Xi_{s+1}$. This implies that

$$\hat{w}_s(\xi) = 1 \qquad \text{on } \operatorname{supp}(\hat{\Psi}_{2s}), \tag{3.4}$$

which will be crucial below. Let w_s denote the inverse Fourier transform of $\hat{w}_s$. Now define

$$f_s = w_s \star f,$$

which is a C^∞ function, so that

$$h_s = 2^{-s}(-\Delta)^{1/4} f_s$$

is well defined. We have the key identity

$$\langle \gamma_\mu, h_s \rangle = \langle c_\mu^+, f \rangle.$$

Moreover, if D_s denotes the bandpass operator introduced in Section 2,

$$\begin{aligned} \|D_s h_s\|_2^2 &= \int |\hat{\Psi}_{2s}(\xi)|^2 |\hat{h}_s(\xi)|^2 \, d\xi \\ &= \int |\hat{\Psi}_{2s}(\xi)|^2 \bigl(2^{-s}|\xi|^{1/2}|\hat{f}(\xi)|\bigr)^2 \, d\xi \\ &= 2^{-2s} \int_{\Xi_s} |\xi| |\hat{\Psi}_{2s}(\xi) \hat{f}(\xi)|^2 \, d\xi, \end{aligned}$$

where in the first step we used (3.4), which gives

$$\hat{w}_s(\xi)\hat{\Psi}_{2s}(\xi) = \hat{\Psi}_{2s}(\xi), \qquad \xi \in \mathbf{R}^2.$$

Now for constants $c_i > 0$ not depending on s,

$$c_1 \min\{|\xi| : |\xi| \in \Xi_s\} \le 2^{2s} \le c_2 \max\{|\xi| : |\xi| \in \Xi_s\},$$

so that with constants not depending on s,

$$2^{-2s} \int_{\Xi_s} |\xi| |\hat{\Psi}_{2s}(\xi) \hat{f}(\xi)|^2 \, d\xi \asymp \int_{\Xi_s} |\hat{\Psi}_{2s}(\xi) \hat{f}(\xi)|^2 \, d\xi.$$

It follows that

$$\|D_s h_s\|_2^2 \asymp \|f_s\|_2^2,$$

with constants not depending on s. Hence, with $\mathcal{M}_s$ the collection of μ with s in the first slot,

$$\begin{aligned} \sum_{\mu \in \mathcal{M}_s} \langle \gamma_\mu^+, f \rangle^2 &= \sum_{\mu \in \mathcal{M}_s} \langle \gamma_\mu, h_s \rangle^2 \\ &= \|D_s h_s\|_2^2 \\ &\asymp \|f_s\|_2^2, \end{aligned}$$

with the next-to-last step following by the tight frame property of γ_μ. Here the constants of equivalence in the last step can be taken independent of s. The result (3.2) now follows by summing across s.

There are two ways to prove the remaining assertion, (3.1). On the one hand, one can proceed concretely, using arguments reminiscent of (3.2). On the other hand, there is an abstract argument. For variety, we take the abstract approach.

It is known in the theory of frames that the stable synthesis property (3.1) (a.k.a. Bessel property) follows from the L^2-norm equivalence (3.2) (see [11] for details). A general result in that theory says that if a general collection (φ_n) of L^2 functions obeys the norm equivalence

$$\sum_n |\langle f, \varphi_n\rangle|^2 \asymp \|f\|_2^2,$$

with positive constants not depending on f, then there exists a dual collection $\tilde{\varphi}_n$ also with the property

$$\sum_n |\langle f, \tilde{\varphi}_n\rangle|^2 \asymp \|f\|_2^2,$$

with implied constants not depending on f, and a reconstruction formula

$$f = \sum_n \langle f, \tilde{\varphi}_n\rangle \varphi_n,$$

with equality holding in the sense of unconditional L^2 convergence of the right-hand side to the left-hand side. Moreover, the dual collection $\tilde{\varphi}_n$ can be chosen so that for any f,

$$\sum_n |\langle f, \tilde{\varphi}_n\rangle|^2 = \min\left\{\sum_n |a_n|^2,\ f = \sum_n a_n \varphi_n\right\}.$$

In a moment we will see that the above equivalences imply

$$\left\|\sum_n a_n \varphi_n\right\|_2 \le C\left(\sum_n |a_n|^2\right)^{1/2}; \tag{3.5}$$

this will complete the abstract argument for obtaining (3.1) from (3.2). Let's see why (3.5) holds. For $f = \sum_n a_n \varphi_n$ we have

$$\begin{aligned}\|f\| &\asymp \left(\sum_n |\langle f, \tilde{\varphi}_n\rangle|^2\right)^{1/2}\\ &\le \min\left\{\left(\sum_n |a_n'|^2\right)^{1/2} : f = \sum_n a_n' \varphi_n\right\}\\ &\le \left(\sum_n |a_n|^2\right)^{1/2}.\end{aligned}$$

The last step follows because, although there may be many ways to synthesize f from appropriate coefficients (a'_n), one particular way is given by the (a_n) (by hypothesis). □

We make a useful remark about the regularity of the $\gamma_\mu^\pm$:

LEMMA 1. *The $\gamma_\mu^\pm$ are smooth and of rapid decay, together with all their derivatives.*

PROOF. By the explicit formula (2.4) we know that in the frequency domain, the $\hat{\gamma}_\mu(\xi)$ are compactly supported in an annular region omitting the origin, and are C^∞. The $\hat{\gamma}_\mu^\pm(\xi)$ are obtained by multiplication with a function that is C^∞ away from the origin. Hence the support of the $\hat{\gamma}_\mu^\pm$ is the same annulus as the corresponding $\hat{\gamma}_\mu$ and on this support, by application of the Leibniz rule, we can obtain bounds on derivatives of all orders. These Fourier-domain conditions imply corresponding space-domain conditions and prove the lemma. □

4. BCD for Radon transform. We now establish the existence of a biorthogonal decomposition of the Radon transform operator driven by the curvelet frame. It is constructed in a fashion reminiscent of the WVD in [18], only using curvelets in place of wavelets.

We begin by recalling to the reader's attention the *Radon isometry*; see Helgason [40].

For a smooth function $f(x) = f(x_1, x_2)$ of rapid decay, let Rf denote the Radon transform of f, the integral along a line $\mathcal{L}_{(\theta,t)}$, expressed using the Dirac mass δ as

$$(Rf)(t,\theta) = \int f(x)\delta(x_1\cos\theta + x_2\sin\theta - t)\,dx, \tag{4.1}$$

where we permit $\theta \in [0, 2\pi)$ and $t \in \mathbb{R}$. For more information about the Radon transform see, for example, [13, 40]. Observe that the line $\mathcal{L}_{(\theta,t)}$ is identical to the line $\mathcal{L}_{(\theta+\pi,-t)}$. As a result, Rf has the *antipodal symmetry*,

$$(Rf)(-t, \theta+\pi) = (Rf)(t,\theta). \tag{4.2}$$

We let $\mathcal{R}$ denote the space of all functions in $L^2(dt\,d\theta)$ with this symmetry.

Define now the operator $\Box^\alpha$ for fractional differentiation of functions of a single variable via

$$(\Box^\alpha f)(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} |\omega|^{\alpha/2}\hat{f}(\omega)\,e^{i\omega t}\,d\omega,$$

with $\Box$ short for $\Box^1$. The *Radon isometry* is then defined by

$$\tilde{R} = (\Box \otimes I)\circ R;$$

it can be well defined on L^2 functions. On all such functions it is an L^2 isometry,

$$[\tilde{R}f, \tilde{R}g] = \langle f, g\rangle,$$

thanks to the Fourier isometry $\langle \hat{f}, \hat{g}\rangle = (2\pi)^2\langle f, g\rangle$. Moreover, $\tilde{R}$ maps the $C^\infty(\mathbf{R}^2)$ functions of rapid decay whose Fourier transforms vanish in the vicinity of the origin into $C^\infty(\mathbf{R} \times [0, 2\pi))$ functions of rapid decay in t.

THEOREM 2. *Define systems* (U_μ) *and* (V_μ) *via* $U_\mu = \tilde{R}\gamma_\mu^+$, $\mu \in \mathcal{M}$ *and* $V_\mu = \tilde{R}\gamma_\mu^-$, $\mu \in \mathcal{M}$. *These are frames for* $\mathcal{R} \subset L^2(dt\, d\theta)$: *the system* (U_μ) *exhibits the almost-orthogonality property*

$$\Bigl\|\sum_\mu a_\mu U_\mu\Bigr\|_2 \le C\Bigl(\sum_\mu a_\mu^2\Bigr)^{1/2} \qquad \forall\, a \in \ell^2 \tag{4.3}$$

and the L^2*-norm equivalence property*

$$\sum_\mu \langle g, U_\mu\rangle^2 \asymp \|g\|^2_{L^2(dt\, d\theta)} \qquad \forall\, g \in \mathcal{R} \subset L^2(dt\, d\theta) \tag{4.4}$$

and similarly for (V_μ). *The two systems are quasi-biorthogonal*,

$$[V_\mu, U_{\mu'}] = 2^{s-s'}\langle \gamma_\mu, \gamma_{\mu'}\rangle, \qquad \mu, \mu' \in \mathcal{M}. \tag{4.5}$$

Put now $\kappa_s = 2^{-s}$. *Then* R *has a biorthogonal decomposition*

$$Rf = \sum_\mu \langle f, \gamma_\mu\rangle \kappa_s V_\mu \tag{4.6}$$

and its adjoint R^* *has decomposition*

$$R^*g = \sum_\mu [g, U_\mu]\kappa_s \gamma_\mu. \tag{4.7}$$

PROOF. Relations (4.3)–(4.5) are applications of Radon isometry. Consider (4.3). The isometry combined with Theorem 1 gives

$$\begin{aligned}\Bigl\|\sum_\mu a_\mu U_\mu\Bigr\|_2 &= \Bigl\|\sum_\mu a_\mu \gamma_\mu^+\Bigr\|_2\\ &\le C\Bigl(\sum_\mu a_\mu^2\Bigr)^{1/2},\end{aligned}$$

establishing the almost-orthogonality (4.3), with a similar argument for (V_μ).

Consider (4.5). Using the isometry property of $\tilde{R}$, we have

$$\begin{aligned}[V_\mu, U_{\mu'}] &= [\tilde{R}\gamma_\mu^-, \tilde{R}\gamma_{\mu'}^+]\\ &= \langle \gamma_\mu^-, \gamma_{\mu'}^+\rangle\\ &= \langle \gamma_\mu, \gamma_{\mu'}\rangle \cdot 2^{s-s'},\end{aligned}$$

which gives (4.5).

Consider finally (4.4). For those g arising as $\tilde{R}f$, with f a finite sum of γ_μ^-'s, we have

$$\begin{aligned}\sum_\mu [U_\mu, g]^2 &= \sum_\mu \biggl[\tilde{R}\gamma_\mu^+, \sum_{\mu'} \alpha_{\mu'} \tilde{R}\gamma_{\mu'}^-\biggr]^2 \\ &= \sum_\mu \biggl\langle \gamma_\mu^+, \sum_{\mu'} \alpha_{\mu'} \gamma_{\mu'}^- \biggr\rangle^2 \\ &\asymp \biggl\| \sum_{\mu'} \alpha_{\mu'} \gamma_{\mu'}^- \biggr\|_{L^2(dx_1\,dx_2)}^2 \\ &= \|g\|_{L^2(dt\,d\theta)}^2,\end{aligned}$$

the first step by Radon isometry, the second step by (3.1), with constants of equivalence not depending on g, and the final step by definition of g and Radon isometry. As $\mathcal{R}$ can be shown to be the L^2 closure of all finite sums of $\tilde{R}\gamma_\mu^-$, (4.4) follows.

An alternative formula for the frames can be given. Start from the well-known *intertwining relation*

$$R \circ (-\Delta)^\alpha = (\Box^{4\alpha} \otimes I) \circ R, \tag{4.8}$$

exhibiting a relationship between fractional powers of the Laplacian in the plane $\mathbf{R}^2$ and fractional differentiation along the t-direction in the Radon domain [40]. Then we have

$$\tilde{R}\gamma_\mu^+ = 2^{-s}(\Box \otimes I)R(-\Delta)^{1/4}\gamma_\mu = 2^{-s}(\Box^2 \otimes I)R\gamma_\mu, \qquad \mu \in \mathcal{M}.$$

In short,

$$U_\mu = 2^{-s}(\Box^2 \otimes I)R\gamma_\mu, \qquad \mu \in \mathcal{M},$$

and similarly

$$\tilde{R}\gamma_\mu^- = 2^s(\Box \otimes I)R(-\Delta)^{-1/4}\gamma_\mu = 2^s R\gamma_\mu, \qquad \mu \in \mathcal{M},$$

so that

$$V_\mu = 2^s R\gamma_\mu, \qquad \mu \in \mathcal{M}.$$

The operator decompositions (4.6) and (4.7) follow immediately. $\Box$

THEOREM 3. *We have the reproducing formula*,

$$f = \sum_\mu [Rf, U_\mu]\kappa_s^{-1}\gamma_\mu, \tag{4.9}$$

valid for all f which are finite sums of γ_μ's.

PROOF. Using the intertwining relation (4.8) from the proof of Theorem 2, one sees immediately that

$$\begin{aligned}[U_\mu, Rf] &= \kappa_s\big[(\Box \otimes I) \circ \tilde{R}\gamma_\mu, Rf\big]\\ &= \kappa_s\big[\tilde{R}\gamma_\mu, (\Box \otimes I) \circ Rf\big]\\ &= \kappa_s\big[\tilde{R}\gamma_\mu, \tilde{R}f\big]\\ &= \kappa_s\langle\gamma_\mu, f\rangle.\end{aligned}$$

In short, *the curvelet coefficients of f are available from the U_μ-based coefficients of Rf*. The relation (4.9) follows immediately. □

In short, f can be obtained from (noiseless, continuous) Radon domain information. However, owing to the $\kappa_s^{-1} = 2^s$ factor this is ill-posed.

LEMMA 2. *The functions $U_\mu(t,\theta)$ and $V_\mu(t,\theta)$ are C^∞ on $\mathbf{R} \times [0, 2\pi)$ and of rapid decay in t.*

For the proof, the argument is essentially the same as Lemma 1's argument for the regularity of $\gamma_\mu^\pm$.

5. Geometry underlying the reproducing formula. The last section built dual frames (U_μ) and (V_μ) for the Radon domain $\mathcal{R} \subset L^2(dt\,d\theta)$. These give a new system of almost-orthogonal analysis and synthesis of "sinograms." They can be viewed as a new set of "features" in analysis of Radon data.

Because of their role in the BCD, they are the Radon-domain features that are the most efficiently analyzed and detected by the methods we develop here. These Radon-domain features are reminiscent of curvelets. They are, at fine scales, highly localized and highly anisotropic.

The BCD associated these features in a one-to-one fashion with curvelets in the original spatial domain, via the reproducing formula (4.9). Conceptually, the reproducing formula performs a kind of edge detection in the Radon domain noting the position and orientation, and the correspondence with curvelets in the space domain allows reconstruction of specific edges at specific positions and orientations in the original domain.

The correspondence between curvelets and dual curvelets as given in the reproducing formula has, at fine scales, an explicit geometric description. Roughly speaking, the curvelet localized near spatial position x_0 and direction θ_0 corresponds to a dual curvelet localized in the Radon plane at (t_0, θ_0) and with direction τ_0, where

$$t_0(x_0, \theta_0) = x_{0,1}\cos(\theta_0) + x_{0,2}\sin(\theta_0) \tag{5.1}$$

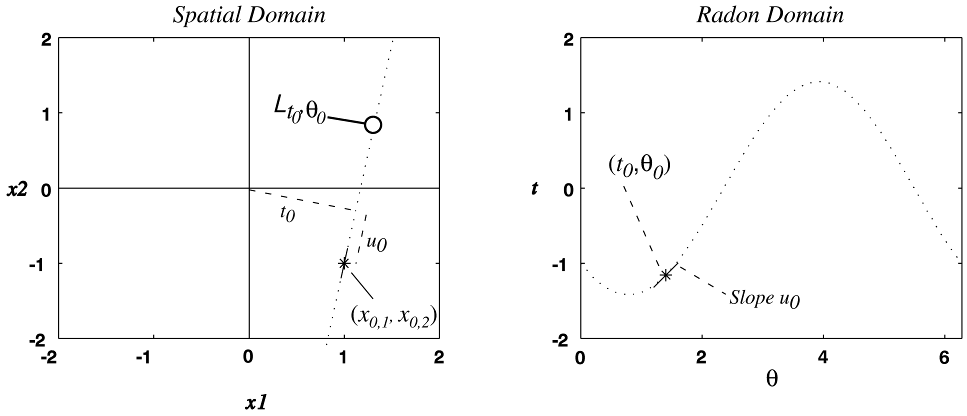

FIG. 1. *Correspondence between spatial domain and Radon domain. A curvelet localized near* $x_0 = (x_{0,1}, x_{0,2})$ *in the spatial domain and oriented in direction* θ_0, *corresponds to a dual curvelet localized near* (t_0, θ_0) *in the Radon domain oriented with slope* u_0. (τ_0 *is the direction, in radians, corresponding to slope* u_0.)

and

$$\tau_0(x_0, \theta_0) = \tan^{-1}[-x_{0,1}\sin(\theta_0) + x_{0,2}\cos(\theta_0)]. \tag{5.2}$$

We illustrate this correspondence in Figure 1.

We state without proof a formal result about this correspondence.

DEFINITION (Joint localization in real space). Suppose we have a sequence (f_n) of functions in $L^2(\mathbf{R}^2)$; we say that the joint location–orientation of f_n converges to (x_0, θ_0) if the space-support of f_n converges to x_0, in the sense that for each $m > 0$,

$$\int |x - x_0|^m |f_n|^2(x)\, dx \to 0 \qquad \text{as } n \to \infty \tag{5.3}$$

and if the direction support of f_n converges to θ_0, in the sense that for each $m > 0$, the Fourier-domain integral

$$\int_0^{2\pi}\int_0^{\infty} \sin(\theta - \theta_0)^{2m} \big|\hat{f}_n(\xi(r,\theta))\big|^2 r\, dr\, d\theta \to 0 \qquad \text{as } n \to \infty. \tag{5.4}$$

DEFINITION (Joint localization in Radon space). Suppose we have a sequence (F_n) of functions in $\mathcal{R} \subset L^2(dt\, d\theta)$; we say that the joint location–orientation of F_n converges to $((t_1, \theta_1), \tau_1)$ if, when we take any smooth window w supported in the vicinity of (t_0, θ_0), and form the windowed extension to a function on $\mathbf{R}^2$,

$$f_n(x_1, x_2) = \begin{cases} w(x_1, x_2) F_n(x_1, x_2), & (x_1, x_2) \in \text{support}(w), \\ 0, & \text{else,} \end{cases}$$

the induced sequence $(f_n) \subset L^2(\mathbf{R}^2)$ has joint location–orientation converging to $x_0 = (t_1, \theta_1)$ and $\theta_0 = \tau_1$ in the sense of the $L^2(\mathbf{R}^2)$ definitions (5.3) and (5.4).

THEOREM 4. *Choose a sequence $\mu_n = (Q_n, \lambda_n)$ of curvelet indices so that*

$$d(Q_n, x_0) \to 0, \qquad n \to \infty$$

and

$$d(2\pi \ell/2^i, \theta_0) \to 0, \qquad n \to \infty.$$

Then:

(a) *the joint location–orientation of γ_{μ_n} converges to (x_0, θ_0);*

(b) *the joint location–orientation of U_{μ_n} converges to $((t_0, \theta_0), \tau_0)$, where t_0, τ_0 are defined in* (5.1) *and* (5.2).

6. Thresholding with noisy data. Suppose now that we observe Radon-domain data according to the white-noise model (1.3). Our goal in this section is to set up some basic terminology and point of view and to explain heuristically the main stages leading to our main result.

6.1. *Analysis of Radon data in white noise.* Assume we have data (1.3). Define empirical coefficients

$$y_\mu = [Y, U_\mu] \equiv \int U_\mu(t, \theta) Y(dt\, d\theta).$$

These obey the Gaussian model

$$y_\mu = [Rf, U_\mu] + \varepsilon[W, U_\mu]. \tag{6.1}$$

Letting α_μ denote the curvelet coefficient $\langle f, \gamma_\mu \rangle$, then from the relation $[Rf, U_\mu] = \kappa_s \langle f, \gamma_\mu \rangle$ we can rewrite (6.1) as

$$y_\mu = \kappa_s \alpha_\mu + \varepsilon n_\mu, \tag{6.2}$$

where κ_s denotes the quasi-singular value, α_μ is the noiseless curvelet coefficient from direct observation of f and $n_\mu \sim N(0, \|\gamma_\mu^+\|_2^2)$ is a (non-i.i.d.) Gaussian noise. In short, using the U_μ system turns a continuum white noise model into a discrete sequence model.

Suppose now that we can construct an estimator $\hat{\alpha} = (\hat{\alpha}_\mu)_\mu$ for the sequence α. Then for the function estimator $\hat{f} = \sum_\mu \hat{\alpha}_\mu \gamma_\mu$ we have, by the tight frame property,

$$E\|\hat{f} - f\|_{L^2(\mathbf{R}^2)}^2 \le E\|\hat{\alpha} - \alpha\|_{\ell^2}^2. \tag{6.3}$$

(This follows from completeness and the fact that $\|\sum_\mu a_\mu \gamma_\mu\|_2^2 \le \sum_\mu |a_\mu|^2$.) In short, estimation error in the curvelet domain controls the estimation error in the original spatial domain. Our strategy is to exploit this fact, and develop estimators in sequence space, knowing that comparable results follow in the continuum model.

This model is similar to models studied in [18, 19, 25] which take the form

$$y_\mu = \kappa_s \alpha_\mu + \varepsilon z_\mu, \tag{6.4}$$

where now z_μ is a standard Gaussian white noise so that noise values z_μ, z_μ are independent for $\mu \neq \mu'$, and the noise is homoscedastic: $\mathrm{Var}(z_\mu) = \mathrm{Var}(z_{\mu'})$ for all μ, μ'. In fact, arguments in [18, 19], combined with (6.3), show that, with the proper recalibration using different ε in each of the two models, results on estimation in the sequence space white-noise model (6.4) yield upper bounds on estimation in the model (6.2). See also the discussion in [44].

6.2. *Thresholding in the white-noise model.* We now briefly mention some existing work on thresholding which will help us to quickly get a rough idea of the mean-squared error properties of thresholding estimators in the white-noise model. The work in [15, 18, 19] suggests one construct estimators in model (6.4) using simple level-dependent thresholding rules.

We begin with a useful heuristic hypothesis. We suppose for the moment that at each level s:

1. There are effectively only M_s coefficients among the $(\alpha_\mu : \mu \in \mathcal{M}_s)$ which can possibly be nonzero. We suppose that $M_s = O(2^s)$. We will discuss the basis of this heuristic in the next subsection.
2. The nonzero coefficients belong to a subset $\mathcal{N}_s$ of $(\alpha_\mu : \mu \in \mathcal{M}_s)$ *which is known a priori* and whose cardinality N_s obeys $N_s = O(2^{4s})$. The basis of this heuristic will also become apparent in a later section. We will refer to N_s as the number of potentially nonzero coefficients.

To summarize, at each level s we have a subset $\mathcal{N}_s$ of N_s coefficients that can potentially be nonzero out of which a maximum of M_s coefficients are effectively nonzero. We do not know the location of the nonzero coefficients ahead of time.

Level-dependent thresholding rules take the form

$$\hat{\alpha}_\mu = \delta(y_\mu \kappa_s^{-1}, t_s), \qquad \mu \in \mathcal{M},$$

where t_s is a level-dependent threshold. A choice which has been well studied in the wavelet setting is to take the threshold at a given level of the expansion to be a certain multiple of the standard error of the underlying statistic, the multiple being determined by the logarithm of the number N_s of potentially nonzero terms at that level of the expansion. This leads to the proposal

$$t_s = \sqrt{2\log(N_s)}\,\kappa_s^{-1}\varepsilon.$$

The oracle inequality [25] gives for $\mu \in \mathcal{N}_s$,

$$E(\hat{\alpha}_\mu - \alpha_\mu)^2 \leq \bigl(2\log(N_s) + 1\bigr)\bigl(\min(\alpha_\mu^2, \kappa_s^{-2}\varepsilon^2) + (\kappa_s^{-2}\varepsilon^2)/N_s\bigr). \tag{6.5}$$

Of course, for $\mu \in \mathcal{M}_s \setminus \mathcal{N}_s$, α_μ is known to be zero and, hence, putting $L_s = (2\log(N_s) + 1)$, the oracle inequality (6.5) gives a total error (across all levels) bounded by

$$E\|\hat{\alpha} - \alpha\|_{\ell^2}^2 \le \sum_s L_s \left(\kappa_s^{-2}\varepsilon^2 + \sum_{\mu \in \mathcal{M}_s} \min(\alpha_\mu^2, \kappa_s^{-2}\varepsilon^2) \right).$$

Ignoring for the moment the logarithmic factor L_s and the term $\varepsilon^2\kappa_s^{-2}$ immediately inside brackets, we focus attention on the expression

$$\sum_s \sum_{\mu \in \mathcal{M}_s} \min(\alpha_\mu^2, \kappa_s^{-2}\varepsilon^2). \tag{6.6}$$

This acts as proxy for the mean-squared error of estimation of a threshold estimator; in studies [25, 19] it has been shown that its behavior mimics, to within logarithmic factors, the true mean-squared error of estimation.

6.3. *Functions that are C^2 away from C^2 edges.* We now formally specify a class of objects with discontinuities along edges; our notation and exposition are taken from [21, 26, 23]; related models were introduced some time ago in the mathematical statistics literature by [47, 48]. It is clear that nothing in the arguments below would depend on the specific assumptions we make here, but the precision allows us to make our arguments uniform over classes of such 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 $\{(1-t)b_0 + tb : t \in [0, 1]\} \subset B$ whenever $b \in B$. This geometrical regularity is useful; it forces very simple interactions of the boundary with dyadic squares at sufficiently fine scales. We use this to guarantee that "sufficiently fine" has a uniform meaning for every B of interest.

We define $\mathrm{S{\scriptsize TAR}}^2(A)$, a class of star-shaped sets with 2-smooth boundaries, by imposing regularity on the boundaries using a kind of polar coordinate system. Let $\rho(\theta) : [0, 2\pi) \to [0, 1]$ 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

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

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

$$\beta(\theta) = \big(\rho(\theta)\cos(\theta) + x_{1,0}, \rho(\theta)\sin(\theta) + x_{2,0}\big). \tag{6.7}$$

Figure 2 gives a graphical indication of some of the objects just described.

The class $\mathrm{S{\scriptsize TAR}}^2(A)$ of interest to us can now be defined by

$$\mathrm{S{\scriptsize TAR}}^2(A) = \Big\{B : B \subset \big[\tfrac{1}{10}, \tfrac{9}{10}\big]^2, \tfrac{1}{10} \le \rho(\theta) \le \tfrac{1}{2},\ \theta \in [0, 2\pi),\ \rho \in \mathrm{H\ddot{O}{\scriptsize LDER}}^2(A)\Big\}.$$

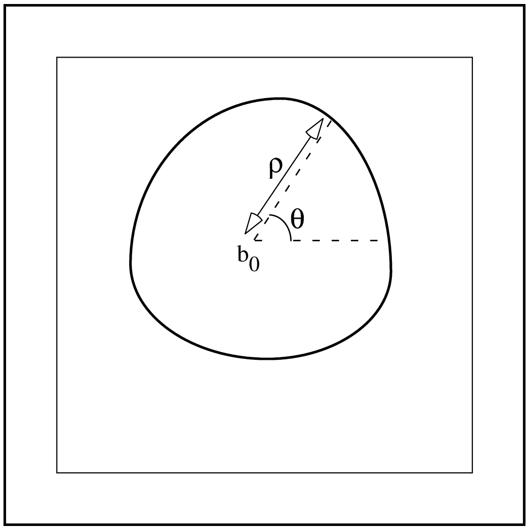

FIG. 2. *Typical star-shaped set and associated notation.*

Here the condition $\rho \in \text{HÖLDER}^2(A)$ means that ρ is continuously differentiable and

$$|\rho'(\theta) - \rho'(\theta')| \le A|\theta - \theta'|, \qquad \theta, \theta' \in [0, 2\pi).$$

The actual objects of interest to us are functions which are twice continuously differentiable except for discontinuities along edges ∂B of star-shaped sets. We define $C_0^2(A)$ to be the collection of twice continuously differentiable functions supported strictly inside $[0, 1]^2$.

DEFINITION. Let $\mathcal{E}^2(A)$ denote the collection of functions f on $\mathbf{R}^2$ which are supported in the square $[0, 1]^2$ and obey

$$f = f_1 + f_2 \cdot \mathbb{1}_B, \tag{6.8}$$

where $B \in \text{STAR}^2(A)$, and each $f_i \in C_0^2(A)$. We speak of $\mathcal{E}^2(A)$ as consisting of FUNCTIONS WHICH ARE C^2 AWAY FROM A C^2 EDGE.

6.4. *The* 4/5 *exponent*: *heuristic argument.* We now consider the risk proxy (6.6) more carefully.

We want to refine now our assumptions about the behavior of the curvelet coefficients.

1. We recall the assumption $M_s = O(2^s)$. The underlying motivation is the fact that a C^2 curve of finite length intersects only $O(2^s)$ dyadic boxes of side 2^{-s}. The assumption is therefore effectively saying that, in forming the curvelet coefficients, at most $O(2^s)$ of the dyadic squares $Q \in \mathcal{Q}_s$ interact with the

discontinuity, each one of those squares has only a few nonzero coefficients and the coefficients from all other squares are trivial at fine scales.

2. We now estimate the size of the coefficients α_μ. In essence, the underlying curvelet frame elements are supported in a region of size 2^{-s} by 2^{-2s}. They are L^2-normalized, so that $\|\gamma_\mu\| \le 1$. They therefore obey

$$\|\gamma_\mu\|_1 \le C2^{-3/2s} \qquad \forall\, \mu.$$

The function f is bounded, so that the nonzero coefficients obey

$$|\alpha_\mu| \le \|\gamma_\mu\|_1 \|f\|_\infty \le C' 2^{-3/2s}.$$

If we now consider a component of the sum (6.6) at one level s, we have

$$\begin{aligned}\sum_{\mu \in \mathcal{M}_s} \min(\alpha_\mu^2, \kappa_s^{-2}\varepsilon^2) &\le M_s \min(C2^{-3s}, 2^{2s}\varepsilon^2)\\ &\le C' 2^s \min(C2^{-3s}, 2^{2s}\varepsilon^2).\end{aligned}$$

Now the right-hand side is largest when s takes the continuum value s^* satisfying

$$C2^{-2s^*} = 2^{3s^*}\varepsilon^2.$$

Simplifying matters so that $C = 1$, we have

$$2^{-s^*} = \varepsilon^{2/5}.$$

Now the *worst level* s^+ will occur at either $\lfloor s^* \rfloor$ or $\lceil s^* \rceil$; at this level

$$\sum_{\mu \in \mathcal{M}_{s+}} \min(\alpha_\mu^2, \kappa_s^{-2}\varepsilon^2) \le C'' \varepsilon^{4/5}.$$

We see that as the level s moves away from s^+, the corresponding sum decays rapidly, so that the sum across levels, (6.6), is bounded by $C''' \varepsilon^{4/5}$.

6.5. *Necessary refinements.* The above arguments give the central organizational ideas which drive our proof of the main result, but there are several ways that they distort the situation.

1. The assumptions about the number $M_s = O(2^s)$ of coefficients that are nonzero which are known to belong to a set $\mathcal{N}_s$, known a priori, of cardinality $N_s = O(2^{4s})$. This is not accurate, since strictly speaking there are a countable number of nonzero coefficients associated with each level of the transform—only most of these are very small compared to any realistic noise level. A valid proof must articulate this fact mathematically and establish the needed approximate notion to replace the assumptions on M_s and $\mathcal{N}_s$.
2. The noise is heteroscedastic, which actually should be exploited, by using a nonuniform threshold within each level.
3. The object is supported in a known region of the plane, which also should be exploited in the thresholding scheme.

7. Main result. We now formally define a curvelet-based reconstruction method and state a result giving the $4/5$ rate of convergence. To obtain a relatively simple proof, we first modify the notion of curvelet expansion. We then deploy this modified expansion in an estimation scheme based on applying threshold to a subset of the noisy dual curvelet coefficients. We then give the basic analysis supporting our theorem. Certain key estimates are relegated to the Appendix.

7.1. *Inhomogeneous curvelet expansions.* The curvelet frame as discussed so far contains elements with support of all sizes, from the very coarse to the very fine. This has been appropriate for developing theoretical decompositions of the Laplacian and the Radon transform, because those operators possess a certain scale-invariance. However, for imaging applications, there is generally a coarsest scale in an image, of size comparable to the largest object in the image. We now introduce a coarsest scale and form an *inhomogeneous curvelet decomposition*.

The inhomogeneous curvelets will be indexed by μ ranging through a set $\mathcal{N}$ which we partition into two sets: $\mathcal{N} = \mathcal{N}^0 \cup \mathcal{N}^1$, corresponding to *coarse* and *fine* scales. For a fixed $s_0 \geq 0$, the fine scale coefficients are indexed by precisely the same scheme as the previous curvelet expansion at all scales finer than s_0,

$$\mathcal{N}^1 = \bigcup_{s \geq s_0} \mathcal{M}_s;$$

the coarse scale coefficients $\mu \in \mathcal{N}^0$ are modeled on the indexing of squares $Q(2s_0, k_1, k_2)$ at scale $2s_0$ and can be thought of as pairs $k = (k_1, k_2)$.

At coarse scales $\mu \in \mathcal{N}^0$, the curvelet coefficients are defined by

$$\alpha_\mu = \langle \phi_{2s_0,k_1,k_2}, P_{2s_0} f \rangle, \qquad \mu \in \mathcal{N}^0,$$

where each ϕ_{2s_0,k_1,k_2} is a Lemarié scaling function [49, 52]. Note that each scaling function is nonoscillatory and that it is localized near a cube $Q(2s_0, k_1, k_2)$.

At fine scales $\mu \in \mathcal{N}^1$, we continue as earlier, with curvelet coefficients simply the multiscale ridgelet coefficients of the filtered object,

$$\alpha_\mu = \langle D_s f, \psi_\mu \rangle, \qquad \mu \in \mathcal{M}_s, \qquad s = s_0, s_0 + 1, \ldots .$$

In short, the inhomogeneous system differs from the earlier homogeneous system only at coarse scales; it differs by collapsing a countable number of scales $s \leq s_0$ into a single scale s_0.

The collection of curvelets γ_μ resulting from our definition has been studied in [8]; it still enjoys the tight frame property.

The arguments of Sections 3 and 4 do not carry through unchanged in the inhomogeneous frame. The coarse scale elements in the inhomogeneous expansion are not oscillatory. As a result, they do not remain in L^2 under fractional integrations. Hence, we cannot apply the full biorthogonal decomposition machinery of Sections 3 and 4 in the inhomogeneous frame. For example, *the companion* γ_μ^- of a coarse-scale element γ_μ *does not exist* as an L^2 function. However, the companion γ_μ^+ *does* exist.

Fortunately, we are still able to construct an estimator. Indeed, for $\mu \in \mathcal{N}^0$ we still have the dual elements U_μ and V_μ and we have the inhomogeneous reproducing formula

$$f = \sum_\mu [Rf, U_\mu]\kappa_\mu^{-1}\gamma_\mu,$$

where the sum is over $\mu \in \mathcal{N}$.

In fact, the exact transcription of Theorem 2 carries through word-for-word in the inhomogeneous frame setting. So working in the inhomogeneous system we still can obtain the curvelet coefficients from an appropriate analysis of the noiseless Radon data.

The only difficulty is that while in the homogeneous setting *all* the U_μ and V_μ were C^∞ and of rapid decay in t, in the inhomogeneous setting the $\mu \in \mathcal{N}^1$ remain C^∞ and of rapid decay, while the $\mu \in \mathcal{N}^0$ remain C^∞ but are now of relatively slow decay in t. In other words, there is no analog of Lemma 2 for the coarse-scale elements $\mu \in \mathcal{N}^0$.

7.2. *Construction of* $\mathcal{N}(\varepsilon)$. Our approach to reconstruction from the noisy data will be to specify a collection $\mathcal{N}(\varepsilon)$ of potentially significant curvelet coefficients, depending on the noise level, and estimate the coefficients in that collection by thresholding, based on statistical significance; all other coefficients will be taken as a priori insignificant, based on an analysis given below, and set to zero.

An important feature of our expansion will be the idea that *the coarsest scale* s_0 *depends on the noise level*. That is, we will actually be varying our inhomogeneous expansion, with the coarsest scale becoming finer and finer as the noise level becomes smaller.

ASSUMPTION (Coarse-scale dependence). *The coarse scale of the inhomogeneous curvelet frame* $s_0 = s_0(\varepsilon)$ *obeys*

$$2^{-s_0-1} \le \varepsilon^{2/15} \le 2^{-s_0}. \tag{7.1}$$

We note that in an asymptotic sense, this coarse scale is still very coarse compared to the scale at which the action mainly takes place. By adopting this notion of scale, we can work with a simply described collection $\mathcal{N}(\varepsilon)$ with simply proved properties.

The underlying reason for our restriction (7.1) is the uncertainty principle, which says that an object cannot be compactly supported in both space and frequency. In this paper, we have elected here to work with bandlimited curvelets, that is, with compact support in frequency. Therefore, at each scale there are infinitely many curvelets overlapping with the support of the object we wish to recover and only a finite number of coefficients can a pirori be potentially significant. In other words, the lack of perfect localization creates the need for bookkeeping to keep track of presumably negligigible behavior in extreme spatial positions; one needs repeatedly to establish decay estimates which verify that behavior at such extremes is indeed negligible. It turns out that, at scales finer than $s_0(\varepsilon)$, the degree of simultaneous localization allows us to greatly simplify certain such bookkeeping arguments.

We would like to emphasize that the restriction (7.1) is merely a choice and is by no means necessary. There are many other ways of defining a collection $\mathcal{N}(\varepsilon)$ of potentially significant curvelet coefficients.

We partition $\mathcal{N}(\varepsilon) = \mathcal{N}^0(\varepsilon) \cup \mathcal{N}^1(\varepsilon)$ into coarse and fine scales, and we start our description by considering coarse scale coefficients. Each coarse-scale $\mu \in \mathcal{N}^0$ is a pair (k_1, k_2).

DEFINITION [$\mathcal{N}^0(\varepsilon)$]. We allow $\mu \in \mathcal{N}^0(\varepsilon)$ if the associated dyadic square $Q = Q(s_0, k_1, k_2)$ interacts with the unit square,

$$d(Q; [0,1]^2) \le 2^{-s_0+1},$$

where $d(\cdot,\cdot)$ is the Hausdorff distance.

At the fine scales $\mu \in \mathcal{N}^1$ indices take the form (Q, λ), where the dyadic square Q is identified by (s, k_1, k_2) with s refering to the scale index and k_1, k_2 to the location of Q, while the parameter $\lambda = (j, k, i, \ell, e)$ indexes the ridgelet transform.

DEFINITION [$\mathcal{N}^1(\varepsilon)$]. We let $\mu \in \mathcal{N}^1(\varepsilon)$ if it obeys

(i) *Localization in scale*. For the scale index s,

$$\tfrac{1}{2}\varepsilon^{2/15} \le 2^{-s} \le \varepsilon^{2/5}. \tag{7.2}$$

(ii) *Localization in space*. For the spatial position Q,

$$d(Q, [0,1]^2) \le 2^{-s+1}.$$

(iii) *Ridgelet localization*. For the ridgelet index $\lambda = (j, k; i, \ell, \varepsilon)$:

(a) *Localization in ridge scale*. The ridge scale parameter j satisfies

$$j = \{s-2, s-1, s, s+1, s+2, s+3, s+4\}.$$

(b) *Localization in angle*. The angular scale parameter i satisfies

$$|i - j| < s.$$

(c) *Localization in ridge location*. The ridge location parameter k satisfies

$$|k| \le 2^{j+1}.$$

7.3. *Properties of $\mathcal{N}(\varepsilon)$.* Underlying our definition of $\mathcal{N}(\varepsilon)$ is the following analysis.

THEOREM 5. (i) *The size of neglected coefficients*,

$$\sup_{f\in\mathcal{E}^2(A)} \sum_{\mu\notin\mathcal{N}(\varepsilon)} |\alpha_\mu|^2 \le C\varepsilon^{4/5}; \tag{7.3}$$

(ii) *the risk proxy*,

$$\sup_{f\in\mathcal{E}^2(A)} \sum_{\mathcal{N}(\varepsilon)} \min(|\alpha_\mu|^2, 2^{2s}\varepsilon^2) \le C\varepsilon^{4/5}; \tag{7.4}$$

(iii) *the number of processed coefficients. The cardinality $N_\varepsilon = \#\mathcal{N}(\varepsilon)$ obeys*

$$N_\varepsilon \le C\varepsilon^{-2}. \tag{7.5}$$

In short,

1. The size of the neglected coefficients obeys a $4/5$ scaling law.
2. The risk proxy obeys a $4/5$ scaling law.
3. The number of estimated coefficients grows polynomially in the inverse noise level.

Each of these properties plays a key role below. Theorem 5 is proved in the Appendix.

7.4. *Definition of estimator.* We now return to the estimation problem (6.1) and (6.2) or equivalently that of estimating the inhomogeneous curvelet coefficients $(\alpha_\mu : \mu \in \mathcal{N})$ from the sequence data

$$y_\mu = \alpha_\mu + \varepsilon\kappa_\mu^{-1} n_\mu,$$

where n_μ is a (non-i.i.d.) Gaussian noise ($E(n_\mu n_{\mu'}) = [U_\mu, U_{\mu'}]$; $\sigma_\mu = \|\gamma_\mu^+\|_2$). We recall that $\kappa_\mu = 2^{-s}$ – ($[Kf, U_\mu] = \kappa_\mu\langle f, \gamma_\mu\rangle$) and that the σ_μ's are uniformly bounded. The same conclusion applies to coarse scale coefficients, $\mu \in \mathcal{N}^0$; namely, $\kappa_\mu = 2^{-s_0}$ and the σ_μ's are uniformly bounded.

We estimate the coefficients α_μ by a thresholding rule and construct an estimator of the form

$$\hat{f} = \sum_{\mathcal{N}(\varepsilon)} \hat{\alpha}_\mu \gamma_\mu.$$

The thresholding is performed as follows. With the soft threshold nonlinearity $\delta(y,t) = \mathrm{sgn}(y)(|y| - t)_+$, we let N_ε be the cardinality of the finite set $\mathcal{N}(\varepsilon)$ and set thresholding parameter $\lambda(\varepsilon) = \varepsilon\sqrt{2\log(N(\varepsilon))}$. We think of this as a "small" multiple of the noise level ε; a statistically significant coefficient will be one which exceeds this. We estimate individual coefficients by the rule

$$\widehat{\alpha_\mu} = \begin{cases} \delta(y_\mu, \lambda\kappa_\mu^{-1}\sigma_\mu), & \mu \in \mathcal{N}(\varepsilon), \\ 0, & \mu \notin \mathcal{N}(\varepsilon). \end{cases} \tag{7.6}$$

In short,

1. The curvelet coefficients α_μ in the "thresholding zone" $\mathcal{N}(\varepsilon)$ are estimated by applying a scalar nonlinearity to the noisy coefficients y_μ; and
2. All the other coefficients are estimated by zero.

7.5. *Analysis of the estimator.* Owing to the tight frame property, we have

$$E\|\hat{f} - \hat{f}\|_2^2 \le E\|\hat{\alpha}_\mu - \alpha_\mu\|_{\ell_2}^2,$$

which allows us to shift attention to the coefficient domain. Considering first the processed coefficients, $\mu \in \mathcal{N}(\varepsilon)$, and applying the oracle inequality (6.5), we have

$$E\sum_{\mathcal{N}(\varepsilon)} (\widehat{\alpha_\mu} - \alpha_\mu)^2 \le \bigl(1 + 2\log(N_\varepsilon)\bigr)\Biggl(\varepsilon^2 \sum_{\mathcal{N}(\varepsilon)} \bigl(\kappa_\mu^{-2}\sigma_\mu^2/N_\varepsilon + \min(\alpha_\mu^2, \varepsilon^2\kappa_\mu^{-2}\sigma_\mu^2)\bigr)\Biggr).$$

Setting $\tau_\mu^2 = \kappa_\mu^{-2}\sigma_\mu^2$, and considering now all $\mu \in \mathcal{N}$, the risk of the thresholding rule (7.6) is hence bounded by

$$E\|\hat{\alpha} - \alpha\|_{\ell_2}^2 \le \bigl(1 + 2\log(N_\varepsilon)\bigr)\Biggl(\varepsilon^2\bar{\tau}^2 + \sum_{\mathcal{N}(\varepsilon)} \min(\alpha_\mu^2, \varepsilon^2\tau_\mu^2)\Biggr) + \sum_{\mathcal{N}(\varepsilon)^c} \alpha_\mu^2, \tag{7.7}$$

where $\bar{\tau}^2$ is simply shorthand for $\{N_\varepsilon\}^{-1}\sum_{\mathcal{N}(\varepsilon)} \tau_\mu^2$. We now estimate systematically each component of this risk bound.

The term on the extreme right-hand side of (7.7) is controlled through Theorem 5's analysis of the size of neglected coefficients; by (7.3) it is $O(\varepsilon^{4/5})$. From that theorem's analysis of the proxy risk (7.4) we have

$$\sum_{\mathcal{N}(\varepsilon)} \min(\alpha_\mu^2, \varepsilon^2\tau_\mu^2) \le C\varepsilon^{4/5}. \tag{7.8}$$

On the other hand, we have

$$\begin{aligned} \varepsilon^2\bar{\tau}^2 &= \varepsilon^2 \sum_{\mathcal{N}(\varepsilon)} \kappa_\mu^{-2}\sigma_\mu^2/N_\varepsilon \\ &\le C\varepsilon^2 \sum_{\mathcal{N}(\varepsilon)} 2^{2s}/N_\varepsilon \\ &\le C\varepsilon^2 \sup_{\mathcal{N}(\varepsilon)} 2^{2s} \qquad (7.9) \\ &\le C\varepsilon^2\varepsilon^{-4/5} \le C\varepsilon^{6/5}. \qquad (7.10) \end{aligned}$$

Here the key step was to invoke scale localization (7.2) at the third display (7.9). Finally, counting the number of estimated coefficients crudely via (7.5) we have

$$\log(N_\varepsilon) = O\big(\log(\varepsilon^{-1})\big). \tag{7.11}$$

Applying now the estimates (7.3), (7.8), (7.10) and (7.11) to (7.7) gives the following upper bound on the risk of the estimator (7.6):

$$\begin{aligned} E\|\hat{\alpha} - \alpha\|_{\ell_2}^2 &\le C\big(\log(\varepsilon^{-1})(\varepsilon^{6/5} + \varepsilon^{4/5}) + \varepsilon^{4/5}\big) \\ &\le C\log(\varepsilon^{-1})\varepsilon^{4/5}. \end{aligned} \tag{7.12}$$

This proves the main result of this paper, the following theorem.

THEOREM 6. *Let $\hat{f}$ be the shrinkage estimator $\hat{f} = \sum_\mu \hat{\alpha}_\mu \gamma_\mu$ where $\hat{\alpha}_\mu$ is given by* (7.6). *Then*

$$\sup_{\mathcal{E}^2(A)} E\|\hat{f} - f\|_2^2 \le C\log(\varepsilon^{-1})\varepsilon^{4/5}. \tag{7.13}$$

7.6. *Variations.* Many variations on the estimation procedure are, of course, possible. For instance, in practice, we may not want to threshold the coarse layer of curvelet coefficients y_μ, $\mu \in \mathcal{N}^0(\varepsilon)$; for example, one might consider estimating those α_μ's with

$$\widehat{\alpha_\mu} = \begin{cases} y_\mu, & \mu \in \mathcal{N}^0(\varepsilon),\ \mu \in K_\varepsilon, \\ 0, & \mu \in \mathcal{N}^0(\varepsilon),\ \mu \notin K_\varepsilon, \end{cases}$$

for some strategic collection of pairs K_ε. Estimators of this kind obey similar bounds.

Because bounds similar to (7.7) exist for hard-thresholding rules, Theorem 6 continues to hold if one replaces the soft-thresholding nonlinearity $\delta(y, \lambda)$ by hard thresholding $\eta(y, \lambda) = y\mathbb{1}_{\{|y| \ge \lambda\}}$, with the same choice of thresholding parameter.

In addition, the authors are confident that further refinements would give versions of Theorem 6 with sharper bounds. In particular, it seems plausible that Hybrid-SURE estimation procedures [44] would allow removing the logarithmic factor from the upper bound (7.13). Such refinements are, however, beyond the scope of the present article.

8. Lower bounds.

The behavior we have established for shrinkage of curvelet coefficients is near-optimal as regards rate of convergence; no estimator can achieve an essentially better rate uniformly over $\mathcal{E}^2(A)$.

THEOREM 7. *Let $\mathcal{E}^2(A)$ be the collection* (6.8) *of objects which are C^2 away from a C^2 curve. The minimax mean-squared error*

$$\mathcal{M}\big(\varepsilon, \mathcal{E}^2(A)\big) = \inf_{\hat{f}} \sup_{\mathcal{E}^2(A)} E\|\hat{f} - f\|_2^2$$

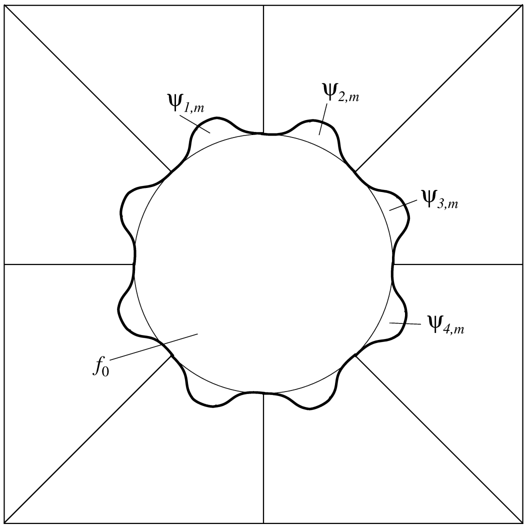

FIG. 3. *The hypercube construction. Here f_0 is the indicator of the circular region*; *around this region are m "bulges"*; *each bulge is disjoint from the other bulges and so the corresponding indicators $\psi_{i,m}$ are orthogonal in $L^2[0,1]^2$. Each combination of the circular region together with a specific collection of bulges makes a specific image which belongs to $\mathcal{E}^2(A)$*; *the 2^m such images can be viewed as a complete m-dimensional hypercube.*

is bounded below by

$$\mathcal{M}\big(\varepsilon, \mathcal{E}^2(A)\big) \geq c\varepsilon^{4/5}(\log \varepsilon^{-1})^{-2/5}, \qquad \varepsilon \to 0.$$

As is standard in lower bounds, the proof relies on the construction of hypercubes embedded in $\mathcal{E}^2(A)$, and then considering the subproblem of estimation when the object comes from one of the vertices of the hypercube. The problem of estimation becomes simply the problem of determining which of the many vertices of the hypercube may have generated the observed data; choosing the hypercube in the appropriate way makes it very difficult to do so. From this lower bound over the hypercube, it follows that no estimator can do very well in the original problem of estimating an unknown f known only to lie in the larger class $\mathcal{E}^2(A)$. The hypercube construction itself is similar to constructions in [21, 23, 26]. See Figure 3.

An important feature of our argument is the fact that, in an inverse problem setting, the problem of inference about f necessarily involves inference about a Normal mean in the presence of correlated noise.

8.1. *Hypercubes.* Let $\varphi(t)$ be a real-valued function of real variable t having compact support $\subset [0, 2\pi]$ with $\|\frac{d^2}{dt^2}\varphi\|_\infty = 1$. With C the constant defining the class $\mathcal{E}^2(A)$, define the collection

$$\varphi_{i,m}(t) = Cm^{-2}\varphi(mt - 2\pi i), \qquad i = 0, 1, \ldots, m-1.$$

Fix an origin at $(1/2, 1/2)$ and define polar coordinates (r, θ) relative to this choice of origin. Set $r_0 = 1/4$ and set $f_0 = \mathbb{1}_{\{r \le r_0\}}$. Consider the collection of functions

$$\psi_{i,m} = \mathbb{1}_{\{r \le \varphi_{i,m} + r_0\}} - f_0.$$

These functions are disjointly supported on lens-shaped regions. They are orthogonal and their common L^2-norm can be calculated approximately; namely,

$$\|\psi_{i,m}\|_2 \asymp C_\varphi m^{-3/2}, \qquad m \to \infty$$

with

$$C_\varphi = \|\varphi\|_{L^1[0,2\pi]}.$$

They generate a hypercube by the prescription

$$\mathcal{H}_m = \left\{ h = f_0 + \sum_i \xi_i \psi_{i,m},\ \xi_i \in \{0, 1\} \right\};$$

each vertex of the hypercube is the indicator of a blob with C^2-smooth boundary, a corrugated disk to which have been appended a certain number of lens-shaped features.

Not only do the vertices $h \in \mathcal{H}_m$ correspond to sets with C^2-smooth boundaries, each one actually obeys the quantitative restriction, $h \in \mathcal{E}^2(A)$. More succinctly, we have the embedding

$$\mathcal{H}_m \subset \mathcal{E}^2(A).$$

Consider the same Radon inversion problem over the restricted function class $\mathcal{H}_m$. We have noisy data

$$Y(dt\, d\theta) = (Rf)(t, \theta)\, dt\, d\theta + \varepsilon W(dt\, d\theta),$$

where now $f \in \mathcal{H}_m$ and so difficulty of estimation is measured by

$$\mathcal{M}(\varepsilon, \mathcal{H}_m) = \inf_{\hat{f}} \sup_{\mathcal{H}_m} E\|\hat{f} - f\|_2^2.$$

From the setwise inclusion $\mathcal{H}_m \subset \mathcal{E}^2(A)$ we have the inequality at the level of minimax risks

$$\mathcal{M}(\varepsilon, \mathcal{E}^2(A)) \ge \mathcal{M}(\varepsilon, \mathcal{H}_m). \tag{8.1}$$

The above argument works for any choice of m, and hence provides a range of lower bounds. Below we will develop lower bounds on the minimax mean-squared error $\mathcal{M}(\varepsilon, \mathcal{H}_m)$, for certain m; these bounds take the form

$$\mathcal{M}(\varepsilon, \mathcal{H}_m) \ge Bm. \tag{8.2}$$

Below, we will specialize to $m = m(\varepsilon)$ satisfying

$$\varepsilon \sim cm^{-5/2}\sqrt{\log(m)} \tag{8.3}$$

for an appropriate constant c. Specializing (8.2) to the choice $m(\varepsilon)$ of (8.3) and invoking (8.1) gives Theorem 7.

In deriving risk bounds on hypercubes, it is useful to note that, on the hypothesis $f \in \mathcal{H}_m$, we can restrict our attention to estimators of the form

$$\hat{f} = f_0 + \sum_i \hat{\xi}_i \psi_{i,m}. \tag{8.4}$$

Indeed, let P_m denote the L^2 projection on the smallest affine subspace containing all such functions. Then since for $f \in \mathcal{H}_m$, $P_m f = f$ we have

$$\|P_m \hat{f} - f\|_2^2 = \|P_m \hat{f} - P_m f\|_2^2 \leq \|\hat{f} - f\|_2^2.$$

Hence the risk of a general estimator $\hat{f}$ is greater or equal to that of a corresponding estimator $P_m \hat{f}$, which of course can be written in the form (8.4).

Moreover, owing to the orthogonality of the $\psi_{i,m}$, we have that for estimators obeying (8.4)

$$\|\hat{f} - f\|_{L^2} = \|\hat{\xi} - \xi\|_{\ell^2}. \tag{8.5}$$

So the problem reduces to one of estimating ξ.

A further reduction is possible. Let now $g_i = R\psi_{i,m}$ denote the Radon-space image of one of our hypercube generators $\psi_{i,m}$. Although the $\psi_{i,m}$ are orthogonal for $L^2(dx_1\,dx_2)$, the g_i are not orthogonal for $L^2(dt\,d\theta)$ in general. However, owing to the invertibility of the Radon transform, the g_i's are linearly independent. Let now V_m denote the affine space

$$Rf_0 + \sum_i \theta_i g_i,$$

for arbitrary choices (θ_i). We note that, for any function $\nu(t, \theta)$ which is $L^2(dt\,d\theta)$ orthogonal to V_m, the law of $\int \nu Y(dt\,d\theta)$ is $N(0, \int \nu^2\,dt\,d\theta)$ independently of ξ. In short, the projection of the Radon data on the span V_m is sufficient for ξ.

Because of the linear independence of the g_j's, the linear functionals $[g_j, f - Rf_0]$ give a nondegenerate set of affine coordinates for $f \in V_m$. Consider now projecting the Radon data onto the g_j's:

$$Y_j = \int g_j Y(dt\,d\theta) - \int g_j Rf_0\,dt\,d\theta.$$

The vector $Y = (Y_j)$ gives a nondegenerate set of affine coordinates for the projection of the Radon data on the space V_m. Hence, the vector $Y = (Y_j)$ is a sufficient statistic for the ξ's and we may restrict our attention to estimators that are (possibly randomized) functions of Y alone. Now

$$Y_j = \sum_i [g_j, g_i]\xi_i + \varepsilon[W, g_j].$$

In matrix notation, $Y \sim N(G\xi, \varepsilon^2 G)$ where G is the Gram matrix of the g_i's; that is, $G_{ij} = [g_i, g_j]$. Because the g_i's are linearly independent, the matrix G is invertible. Then define $X = G^{-1}Y$. As Y is a sufficient statistic for the ξ's, so is X and, hence, we may restrict our attention to estimators that are (possibly randomized) functions of X alone.

Because of the risk isometry (8.5) the problem becomes to estimate, under squared ℓ^2-norm loss, the mean $\xi \in \{0,1\}^m$ of a multivariate Gaussian vector from $X \sim N(\xi, \varepsilon^2 G^{-1})$.

The lemma in Section 8.2 gives a lower bound on the minimax risk for estimating ξ, of the form

$$\inf_{\hat{\xi}} \sup_{\xi \in \{0,1\}^m} E\|\hat{\xi} - \xi\|_2^2 \geq Bm,$$

with B an absolute constant. The condition for applying that lemma is that the "noise level" in each coordinate is at least one, that is, each conditional Gaussian law $\mathcal{L}(X_i | (X_j : j \neq i))$ has variance at least 1.

To check this condition, let V be the covariance matrix of X, $V = \varepsilon^2 G^{-1}$ and τ_i^2 be the conditional variance of X_i given the other coordinates

$$\tau_i^2 = \mathrm{Var}(X_i | X_j,\ j \neq i).$$

With this notation, we have

$$1/\tau_i^2 = (V^{-1})_{ii} = \varepsilon^{-2} G_{ii} = \varepsilon^{-2}\|g_i\|^2 = \varepsilon^{-2}\kappa_m^2, \quad \text{say},$$

where $\kappa_m = \|g_i\|_{L^2(dt\,d\theta)}$ is studied in Section 8.3.

Now if we take the smallest m so that

$$\varepsilon^{-2}\kappa_m^2 \leq 1, \tag{8.6}$$

then we (just barely) achieve the desired noisiness in the conditional laws of X: $\mathrm{Var}(X_i | X_j,\ j \neq i) \geq 1$. From Section 8.3, we have the asymptotic relation

$$\kappa_m \leq c m^{-5/2}\sqrt{\log(m)}, \qquad m \to \infty,$$

with c an absolute constant. From this we derive that an $m(\varepsilon)$ obeying (8.3) gives the required noisiness.

8.2. *Bayes risk on hypercubes, dependent data.*

LEMMA 3. *Let π be the prior on $\theta \in \{0,1\}$ which puts equal probability on both outcomes and let B be the Bayes risk of estimating θ from $X \sim N(\theta, 1)$. Observe $Y \sim N(\theta, V)$, $\theta \in \{0,1\}^n$, and suppose that $\tau_i^2 = 1$. Then*

$$\inf_{\hat{\theta}} \sup_{\theta \in \{0,1\}^n} E\|\hat{\theta} - \theta\|_2^2 \geq Bn.$$

PROOF. The proof of the lower bound follows an argument developed in [44]. The indicated minimax risk exceeds the Bayes risk of any particular choice of prior π on θ,

$$\inf_{\hat{\theta}} \sup_{\theta \in \{0,1\}^n} E\|\hat{\theta} - \theta\|_2^2 \geq \inf_{\hat{\theta}} E_\pi E\|\hat{\theta} - \theta\|_2^2.$$

Consider then the prior defined by setting the components θ_i to be independent, and $P(\theta_i = 0) = P(\theta_i = 1) = 1/2$; we obtain a lower bound by calculating its Bayes risk. (This is not necessarily the least-favorable prior, but this choice produces lower bounds of the correct order.)

Let us introduce the random variables ξ_i defined by

$$\xi_i = \theta_i + \tau_i^2 \{V^{-1}(Y - \theta)\}_i.$$

From $Y - \theta \sim N(0, V)$, one easily deduces $V^{-1}(Y - \theta) \sim N(0, V^{-1})$ and, therefore,

$$\xi_i - \theta_i \sim \tau_i^2 N\bigl(0, (V^{-1})_{ii}\bigr) = \tau_i^2 N(0, 1/\tau_i^2) = N(0, \tau_i^2).$$

In [44], the authors argue that the distribution of θ_i conditional on $\{Y, \theta_j \text{ for } j \neq i\}$ is the same as that conditional on ξ_i. This is useful to bound the Bayes risk $B(\pi)$ from below:

$$\begin{aligned} B(\pi) &= \sum_i E(\hat{\theta}_i - \theta_i)^2 \\ &= \sum_i E\bigl(\mathrm{Var}(\theta_i | Y)\bigr) \\ &\geq \sum_i E\bigl(\mathrm{Var}(\theta_i | Y, \theta_j \text{ for } j \neq i)\bigr) \\ &= \sum_i E\bigl(\mathrm{Var}(\theta_i | \xi_i)\bigr). \end{aligned}$$

The result now follows from the definition of B. □

8.3. *Norm estimates.*

LEMMA 4. *For $\kappa_m = \|g_i\|$,*

$$\kappa_m \asymp m^{-5/2}\bigl(\log(m)\bigr)^{1/2}, \qquad m \to \infty.$$

PROOF. We begin the proof by observing that, for any $\psi_{i,m}$, there are ellipses $\mathcal{E}$ and $\mathcal{E}'$, both with major and minor axes of sidelength $\sim m^{-1}$ and $\sim m^{-2}$, respectively, such that

$$\mathbb{1}_{\mathcal{E}} \leq \psi_{i,m} \leq \mathbb{1}_{\mathcal{E}'}.$$

Then the Radon transforms satisfy

$$R\{\mathbb{1}_{\mathcal{E}}\} \le g_i \le R\{\mathbb{1}_{\mathcal{E}'}\}.$$

Moreover, because of the geometric similarity of the support of $\psi_{i,m}$ to the fixed lens-shaped region

$$\{(y, \varphi(y)) : 0 \le y \le 2\pi\},$$

we can arrange that the ratio of axis lengths for the inscribing and circumscribing ellipses stays bounded independently of both i and m. We now calculate the size of the bracketing Radon transforms.

We first remark that the Radon transform has the following natural covariance property. Let U be an orthogonal matrix representing planar rotation by θ_0 radians and let $b \in \mathbf{R}^2$. Then

$$R\{f(Ux - b)\}(t, \theta) = R\{f(x)\}(t - U^T b, \theta - \theta_0). \tag{8.7}$$

Using this, we reduce consideration to a standard ellipse $\mathcal{E} = \{x^2/a^2 + y^2/b^2 \le 1\}$. We get

$$R\{\mathbb{1}_{\mathcal{E}}\}(t, \theta) = ab/\sigma(1 - t^2/\sigma^2)^{1/2}, \qquad \sigma^2 = a^2\cos^2\theta + b^2\sin^2\theta.$$

This follows from the fact that the Radon transform of the unit disk $g(x, y) = \mathbb{1}_{\{x^2+y^2 \le 1\}}$ is given by

$$Rg(t, \theta) = 2(1 - t^2)^{1/2}$$

and a simple change of variable.

Now, take $a = m^{-2}$, $b = m^{-1}$ and let us calculate the L^2-norm of $R\{\mathbb{1}_{\mathcal{E}}\}$. We have

$$\begin{aligned} \|R\{\mathbb{1}_{\mathcal{E}}\}\|_2^2 &= \int_0^{2\pi} \int_{-\sigma_m}^{\sigma_m} (m^{-3}/\sigma_m)^2 (1 - t^2/\sigma_m^2)\, dt\, d\theta \\ &= m^{-6} \int_0^{2\pi} \sigma_m^{-1}\, d\theta \int_{-1}^{1} (1 - u^2)\, du \\ &\sim m^{-5} \int_0^{2\pi} (m^{-2}\cos^2\theta + \sin^2\theta)^{-1/2}\, d\theta \\ &\sim m^{-5} \log m. \end{aligned} \tag{8.8}$$

The desired conclusion follows from (8.8) and the invariance (8.7). □

9. Discussion.

9.1. *Deconvolution.* So far, we have focused entirely on Radon inversion. However, our approach is more general, able to give results for a range of inverse problems. We briefly mention results for an inverse problem of deconvolution which follow as corollaries of the results above.

Define the two-dimensional *Bessel potential* of order α, the kernel $b_\alpha(x)$ with Fourier transform

$$\hat{b}_\alpha(\xi) = (1 + |\xi|^2)^{-\alpha/2}, \qquad \xi \in \mathbf{R}^2.$$

The Bessel operator B_α is the operator of convolution with b_α: $B_\alpha = b_\alpha \star f$. Below we will study an inverse problem based on this operator, but first we consider the operator-biorthogonal curvelet decomposition of B_α. Recalling the general definition given in Section 1.7, and applying it now in the case $K = B_\alpha$, we can see that all the required machinery works smoothly, producing a decomposition with $\kappa_s = 2^{-2\alpha s}$ for all $s \geq s_0$.

The key insight is that the Bessel potential operator has well-known connections to the Riesz fractional integration operator

$$I_\alpha(f)(x) = c_\alpha \int \frac{f(y)}{\|x - y\|^{2-\alpha}}\, dy,$$

where $c_\alpha = \Gamma(1-\alpha/2)/(\pi 2^\alpha \Gamma(\alpha/2))$; compare [59]. In particular, the two kernels behave comparably near the origin,

$$b_\alpha(x) \sim c_\alpha |x|^{\alpha-2}, \qquad |x| \to 0.$$

More to the point, B_α behaves, at fine scales, like fractional integration of order α, in the sense that, when applied to an oscillatory function $g_{a,b}(x) = g((x-b)/a)$ with small a, we have

$$B_\alpha(g_{a,b}) \approx I_\alpha(g_{a,b})$$

in various senses. Put another way, the frequency response of the fractional integration operator I_α is exactly $|\xi|^{-\alpha}$ while the frequency response of the Bessel potential operator is $(1 + |\xi|^2)^{-\alpha/2}$, evidently asymptotic to $|\xi|^{-\alpha}$ as $|\xi| \to \infty$. The Bessel operator of order α principally differs from fractional integration of order α at low frequencies, where it is far better behaved. This key distinction is responsible for the fact that b_α is of rapid decay as $|x| \to \infty$, while the Riesz kernel is of slow decay.

Now the fractional integration operator is precisely a (negative) fractional power of the Laplacian: on nice functions f,

$$I_\alpha(f) = (-\Delta)^{\alpha/2} f.$$

We recall the results of Section 3, which showed that curvelets provide a biorthogonal decomposition of $\Delta^{-\alpha}$. Proceeding analogously, we can construct a pair of frames $(\gamma_\mu^\sharp)$ and $(\gamma_\mu^\flat)$ which furnish a biorthogonal decomposition of the Bessel operator B_α. Moreover, this works equally as well for the inhomogeneous curvelet decomposition as for the homogeneous one. In what follows, we assume the inhomogeneous curvelet expansion.

Consider the two-dimensional inverse problem in white noise,

$$Y(dt) = (B_\alpha f)(t)\,dt + \varepsilon W(dt), \qquad t \in \mathbf{R}^2, \tag{9.1}$$

where, as in (1.3), f is a compactly supported function which is C^2 away from a C^2 edge. In effect, the Bessel transform smoothes out the edges and otherwise blurs the object.

Using the operator-biorthogonal decomposition of the Bessel kernel, we can propose in the setting (9.1) to obtain noisy curvelet coefficients

$$y_\mu = \kappa_s^{-1}\langle Y, \gamma_\mu^\sharp\rangle,$$

and then to reconstruct f by the thresholding rule

$$\hat{f} = \sum_{\mu \in \mathcal{N}_\alpha(\varepsilon)} \delta(y_\mu, t_\varepsilon \kappa_s^{-1})\gamma_\mu. \tag{9.2}$$

For an arbitrary value of $\alpha > 0$, one would use, here, a slightly different version of our definition of "important coefficients" $\mathcal{N}_\alpha(\varepsilon)$ introduced in Section 7.2 (see below for details) and the same threshold t_ε, defined in the Radon case.

The reuse of the Radon concepts makes sense in the Bessel setting, at least if $\alpha = 1/2$. The Gram operator of the Radon transform is simply a fractional power of the Laplacian; in fact,

$$R^* R = I_1.$$

On the other hand, the Gram operator of the Bessel potential of order $1/2$ is precisely

$$B_{1/2} B_{1/2} = B_1.$$

Our earlier discussion supports a close quantitative similarity of the two operators I_1 and B_1 at fine scales. This similarity can also be seen as follows. We have already seen, in Section 4, that there is a precise formal correspondence between fractional integration of order $1/2$ and Radon transform,

$$\langle (-\Delta)^{-1/4}\gamma_\mu, \gamma_{\mu'}^+\rangle = \kappa_s\langle \gamma_\mu^-, \gamma_{\mu'}^+\rangle = \kappa_s[V_\mu, U_{\mu'}] = [R\gamma_\mu, U_{\mu'}].$$

At the same time, it is clear that, because of the similarity of $(-\Delta)^{-1/4}$ and the Bessel potential of order $1/2$ at high frequencies, we have the approximate relation

$$\langle (-\Delta)^{-1/4}\gamma_\mu, \gamma_{\mu'}^+\rangle \approx [B_{1/2}(\gamma_\mu), \gamma_{\mu'}^\sharp]$$

in various senses, where the approximation improves at successively finer scales.

Recall that in our construction of $\mathcal{N}(\varepsilon)$ we consider only scales finer than a certain cutoff $s_0(\varepsilon)$, and we let $s_0(\varepsilon) \to \infty$ as $\varepsilon \to 0$. Notice that there is very nearly an isometry between tail sections of the $\gamma^{\sharp}_{\mu'}$ system and of the $\gamma^{+}_{\mu'}$ system. It is apparent that, when $\alpha = 1/2$, all relevant properties in the analysis of the estimator (9.2) will be equivalent, to within bounded factors, to corresponding properties derived in the Radon setting.

It is clear that all the quantitative asymptotics that were developed to study the Radon case apply immediately to the estimator (9.2) in the case $\alpha = 1/2$. In short, we have:

1. A $4/5$ law for the curvelet estimator. Suppose that $f \in \mathcal{E}^2(A)$, and we have to recover f from noisy Bessel data (9.1). For each $\delta > 0$, the estimator (9.2) obeys

$$E\|\hat{f} - f\|_2^2 \le C\varepsilon^{4/5+\delta}, \qquad \varepsilon \to 0,$$

where C is the same for all $f \in \mathcal{E}^2(A)$.
2. A $4/5$ law for the lower bound. Precisely the same arguments used to establish Theorem 7 will establish that, for all measurable functions of the observations (9.1) and all $\delta > 0$, we have

$$\sup_{\mathcal{E}^2(A)} E\|\hat{f} - f\|_2^2 \ge c_\delta \varepsilon^{4/5-\delta}, \qquad \varepsilon \to 0.$$

3. A $2/3$ law for wavelet-based approaches. The best thresholding estimator based on wavelet–vaguelette decomposition obeys

$$\sup_{\mathcal{E}^2(A)} E\|\hat{f} - f\|_2^2 \ge c\varepsilon^{2/3}, \qquad \varepsilon \to 0.$$

4. A $1/2$ law for linear deconvolution approaches. The best linear estimator for deconvolution obeys

$$\sup_{\mathcal{E}^2(A)} E\|\hat{f} - f\|_2^2 \ge c\varepsilon^{1/2}, \qquad \varepsilon \to 0.$$

We summarize formally in the corollary.

COROLLARY 1. *Consider the inverse problem of recovering an object $f \in \mathcal{E}^2(A)$ from noisy blurred data* (9.1), *where $\alpha = 1/2$. The method* (9.2) *achieves essentially the optimal rate $\varepsilon^{4/5}$ throughout $\mathcal{E}^2(A)$, outperforming wavelets* (*which achieve only the $\varepsilon^{2/3}$ rate*) *and linear methods* (*which achieve only the $\varepsilon^{1/2}$ rate*).

Corresponding results can be expected to hold for other deconvolution problems with different α. Of course such results would hold with different exponents than $4/5$, $2/3$ and $1/2$. However, we suspect that an α-dependent coefficient set $\mathcal{N}_\alpha(\varepsilon)$ would be required for construction of an effective estimator. For instance, if one defines the set of "important coefficients" $\mathcal{N}_\alpha(\varepsilon)$ as in Section 7.2 with the sole modification that the scale index s is now restricted to the range

$$\tfrac{1}{2}\varepsilon^{2/15} \le 2^{-s} \le \varepsilon^{1/(3/2+2\alpha)}, \quad \text{say.}$$

Compare (7.2); then preliminary calculations show that our curvelet estimator would achieve an estimation rate with exponent $2/(3/2+2\alpha)$.

Exploring this more general situation seems an interesting project for further research.

9.2. *Generalization.* The possibility of generalizing to deconvolution problems is no accident. In fact there is a general strategy for treating inverse problems, of which this article gives an example. The strategy can be formulated as a slogan:

> We seek to construct decompositions for the object and data domains which *almost diagonalize* the Gram operator K^*K of K and which *almost optimally sparsify* the typical object to be recovered. We then exploit the new representations to address inverse problems with noisy data.

In this setting, the Gram operator is a fractional power of the Laplacian and the almost diagonality of the Gram operator is expressed by the frame bound results in Sections 3 and 4. The optimal sparsity is expressed by the effectiveness of curvelets at representing the object of interest with very few coefficients.

There are other examples of this high-level strategy at work: first, with wavelets [18] and then with mirror wavelets [46]. We believe that many other examples are possible.

9.3. *A challenge.* As indicated in the introduction, there is an extensive literature on edge-preserving smoothing and edge-preserving deconvolution. In fact the literature falling in these two categories is far too extensive for us to give a satisfactory set of representative citations.

Much of the literature on this topic is purely methodological. Typical articles in that literature construct computational methods which seem, on general grounds, appropriate, and which exhibit numerous examples of subjectively "successful" reconstructions in specific cases.

Here we have pursued a different strategy, of developing a theoretical model and an optimality result for that model. To our knowledge, this is an innovation in the area of edge preserving denoising–deconvolution.

We have heard, in conferences, oral presentations in which claims have been aired to the effect that certain specific image processing algorithms offer "an optimal way of processing images." For example, such claims have been given in oral presentations in connection with total-variation based image processing methods. Such claims would require the development of statistical–mathematical models similar to ours and some careful analyses establishing some kinds of statistical–mathematical optimality. We are not aware, however, of efforts in this direction. Moreover, we believe that in the model we are considering, such optimality claims for preexisting methods would be false. For example, we believe that in the setting of Radon inversion discussed here, the method of regularized inversion based on total-variation penalization, properly translated into this setting, would not achieve the optimal rate $4/5$.

Our results pose an implicit challenge to all the existing methodological work, which we now make explicit.

CHALLENGE. Prove or disprove that existing methods of edge-preserving recovery achieve or do not achieve the optimal rates we have identified in this article.

This would require a major initiative, subjecting a large body of methodological efforts at edge-preserving reconstruction to a mathematical performance standard.

This initiative is important because, although the major developers of edge-preserving technology may project a great deal of confidence in their tools, we believe that the full story may be considerably different than they imagine. We know of no evidence to suggest that existing proposals for edge-preserving methods achieve optimal rates, and we believe there is good reason to believe that they do not. In particular, we don't believe that total variation penalization can achieve the optimal rate $4/5$ in this setting; instead we consider it likely that it achieves at best the $2/3$ rate of wavelet-based methods in this setting.

Our beliefs are based on mathematical structures underlying the main results of this paper. The method proposed here deploys a system of anisotropic elements to achieve a certain performance upper bound. The matching lower bound also employs anisotropic elements. We believe that the appearance of similar anisotropic features in both lower bounds and upper bounds points to a fundamental correctness or well adaptedness of the approach.

In common sense terms, we have shown that one can reconstruct an edge accurately using a very anisotropic smoothing mechanism, analogous to having an elongated kernel oriented precisely along the edge and having dimensions scaling like ℓ by ℓ^2. We have shown that a very challenging case for reconstruction is to build a family of images, all depicting the indicator of a set, and all differing from each other by the appending or excising of certain elongated regions of size ℓ by ℓ^2.

This suggests to us the strong possibility that a method achieving the optimal rate must be based on a particular kind of adaptive anisotropic smoothing. Indeed, elementary calculations show that if we simply consider the special system constructed for the lower bound, and consider spatially variable kernel methods properly aligned with the edge but not scaled according to the *width* $=$ *length*^{2} principle, for example obeying *width* $=$ *length*, the performance achieved will not scale as $4/5$.

In the existing literature on edge-preserving reconstruction we have never encountered anisotropic smoothing consistent with the scaling law *width* $=$ *length*^{2}. Instead, when anisotropy is invoked, it is largely the very weak kind of anisotropy where one axis of a filter kernel is slightly amplified in length compared to the other. From the viewpoint of the asymptotics we discuss here, such anisotropy is rather weak, and cannot substantially improve rates of convergence.

Given the great deal of interest in nonlinear edge-preserving image processing methods, the issue raised above would seem to be a vital next question.

APPENDIX: PROOF OF THEOREM 5

We collect here various estimates which establish the conclusions of Theorem 5.

A.1. Size of neglected coefficients. We begin by establishing (7.3). In the subsections below, we establish three inequalities, (A.1), (A.5) and (A.7), each of which bounds the sum of squares of the neglected coefficients in a certain subset of $\mathcal{N}(\varepsilon)^c$. Each of these bounds is uniform over $\mathcal{E}^2(A)$ and is of size $C\varepsilon^{4/5}$. The three subsets combine to cover $\mathcal{N}(\varepsilon)$ completely. Hence we conclude

$$\sup_{f\in\mathcal{E}^2(A)} \sum_{\mu\notin\mathcal{N}(\varepsilon)} |\alpha_\mu|^2 \le C\varepsilon^{4/5},$$

which is (7.3).

A.1.1. *Localization in scale.* We now show that, with a fine scale cutoff s_ε satisfying $2^{-s_\varepsilon} \le \varepsilon^{-2/5}$, the sum of squares at all finer scales obeys

$$\sum_{s>s_\varepsilon} \sum_{\mu\in M_s} |\alpha_\mu|^2 \le C\varepsilon^{4/5}. \tag{A.1}$$

The sum in question can be reexpressed according to

$$\sum_{s>s_\varepsilon} \sum_{\mu\in M_s} |\alpha_\mu|^2 = \sum_{s>s_\varepsilon} \|D_s f\|_2^2, \tag{A.2}$$

where D_s is the passband filtering operator. The lemma immediately below implies that this last sum is bounded by $C2^{-2s_\varepsilon}$, and since $2^{-s_\varepsilon} \le \varepsilon^{2/5}$, this yields (A.1). It remains to state and prove the lemma.

LEMMA 5.

$$\sup_{f\in\mathcal{E}^2(A)} \|D_s f\|_2^2 \le C2^{-2s}. \tag{A.3}$$

PROOF. Consider a standard two-dimensional wavelet basis $\phi_{j,k_1,k_2,e}$ using smooth wavelets of compact support and three vanishing moments. We will estimate the number and size of coefficients $\langle f, \phi_{j,k_1,k_2,e}\rangle$ at level j. For each wavelet at scale j which intersects the support of the edge curve, the coefficient obeys an amplitude bound $a_{0,j} = C2^{-j}$. There are at most $n_{0,j} = C2^j$ such coefficients. For every wavelet which does not intersect the edge curve, the coefficient obeys an amplitude bound $a_{1,j} = C2^{-3j}$ and there are at most $n_{1,j} = C2^{2j}$ such coefficients which intersect the support of f at all. We conclude that the sum of squares of the wavelet coefficients at level j is at most

$$\sum_{k_1,k_2,e} \langle f, \phi_{j,k_1,k_2,e}\rangle^2 \le n_{0,j} A_{0,j}^2 + n_{1,j} a_{1,j}^2 \le C2^{-j}, \qquad j \ge j_0.$$

We note that this is uniform over all members $f \in \mathcal{E}^2(A)$, because of the uniform control that such membership brings on the size of f and its derivatives, and also the length of the edge curve.

Now a standard principle of Littlewood–Paley analysis [37, 52] is that the sum of squares of wavelet coefficients at level j is equivalent, within fixed constants, to the squared L^2 norm of a single-octave bandpass filter with passband centered at frequency 2^j.

Now D_s is a double-octave passband filter with passband centered at frequency 2^{2s}. We conclude that the squared L^2 norm of $D_s f$ is at most a constant times 2^{-2s}. We note that, because the estimate on the norm of the wavelet coefficients was uniform over all $f \in \mathcal{E}^2(A)$, so is the estimate on the passband norm. Then (A.3) follows. □

A.1.2. *Localization in space.* We now show that at scales coarser than the fine-scale cutoff, we may neglect the squares Q separated from the support cube of f, $Q_0 = [0, 1]^2$, with total neglected coefficient energy bounded by $C\varepsilon^{4/5}$.

Recall that the curvelet coefficient α_μ at $\mu = (Q, \lambda)$ is given by

$$\alpha_\mu = \alpha_{Q,\lambda} = \langle w_Q(D_s f), \rho_{Q,\lambda}\rangle,$$

where D_s is the operator of convolution by $2^{4s}\Psi(2^{2s}\cdot)$. If the bandpass kernel Ψ were compactly supported then $w_Q(D_s f)$ would be identically zero for any Q such that $d(Q, Q_0) \ge C2^{-s}$. In that case, all curvelet coefficients associated to such squares would vanish and there would be really nothing to prove.

However, in our definition of the curvelet transform, we chose the bandpass filter kernel Ψ to be compactly supported in frequency (and therefore not in space). As a consequence, $D_s f$ is in general not compactly supported but only rapidly decaying away from the unit square. We must therefore estimate the energy content of such a decaying object in squares which are not neighboring to Q_0. The central point is that the squares are of side 2^{-s} while the decay is happening on the scale 2^{-2s}. Therefore at fine scales, by the time the function is at least one square away from Q_0, it is extraordinarily small.

The following lemma is well known in classical analysis; we use it below and prove it here for the convenience of statisticians and perhaps others.

LEMMA 6. *Let $g : \mathbf{R}^n \to \mathbf{R}$ be an arbitrary function such that for some $m > 1$, we have*

$$|g(x)| \leq C_m(1 + |x|)^{-m},$$

for some constant C_m and let $g_{a,b}$ be the dilated translation of g defined by $g((x - b)/a)$. Then, there is a constant C'_m such that

$$\left|\int_{Q_0} f(x) g_{a,b}(x)\, dx\right|^2 \leq C'_m a^{-1}\big(1 + ad(b, Q_0)\big)^{-2m+1} \|f\|_2^2. \tag{A.4}$$

PROOF. Letting $I_{a,b} \equiv \int_{Q_0} f(x) g_{a,b}(x)\, dx$, we have

$$\begin{aligned} |I_{a,b}| &\leq \int_{Q_0} |f(y)| C_m (1 + a|y - b|)^{-m}\, dy \\ &\leq \left(\int_{Q_0} |f(y)|^2\, dy\right)^{1/2} \left(\int_{Q_0} C_m^2 (1 + a|y - b|)^{-2m}\, dy\right)^{1/2}. \end{aligned}$$

Now since $|y - b| \geq \max_i |y_i - b_i|$ we have

$$(1 + a|y - b|)^{-2m} \leq \min_i (1 + a|y_i - b_i|)^{-2m}.$$

Hence for $i = 1, \ldots, n$,

$$\begin{aligned} \int_{Q_0} (1 + a|y - b|)^{-2m}\, dy &\leq \int_0^1 (1 + a|y_i - b_i|)^{-2m}\, dy_i \\ &\leq Ca^{-1}\big(1 + ad(b_i, [0, 1])\big)^{-2m+1}. \end{aligned}$$

On the other hand,

$$\max_i d(b_i, [0, 1]) \geq d(b, Q_0)/\sqrt{n},$$

which implies

$$\int_{Q_0} (1 + a|y - b|)^{-2m} \leq Ca^{-1}\big(1 + ad(b, Q_0)\big)^{-2m+1}.$$

Then (A.4) follows. □

The rapid decay of Ψ implies that for each $m \geq 0$, there is a constant C_m such that

$$|\Psi(x)| \leq C_m(1 + |x|)^{-m},$$

and obviously $|\Psi_{2s}(x)| \leq C_m 2^{4s}(1+2^{2s}|x|)^{-m}$. From

$$(D_s f)(x) = \int_{Q_0} f(y)\Psi_{2s}(x-y)\,dy,$$

we see that Lemma 6 gives the existence of a constant C so that

$$|D_s f|(x) \leq C 2^{3s}\big(1+2^{2s}\,d(x, Q_0)\big)^{-m+1/2}\|f\|_2.$$

It follows that on $x \in Q$,

$$|D_s f|(x) \leq C 2^{3s}\big(1+2^{2s}\,d(Q, Q_0)\big)^{-m+1/2}\|f\|_2,$$

and so, from $\|w_Q\|_{L^2} \leq C2^{-s}$,

$$\|w_Q D_s f\|_2^2 \leq C 2^{4s}\big(1+2^{2s}\,d(Q, Q_0)\big)^{-2m+1}\|f\|_2^2.$$

This bound, which holds with a constant C depending only on m, expresses the exceptionally fast decay of the size of $D_s f$ as Q moves away from Q_0.

Let now $\mathcal{Q}_s'$ denote the collection of squares at scale s obeying $d(Q, Q_0) \geq 2\cdot 2^{-s}$. Then because the ridgelets ρ_λ are orthonormal, we can calculate the norm of the corresponding curvelets associated to a square Q from the norm of the object $w_Q D_s f$,

$$\sum_\lambda |\alpha_{Q,\lambda}|^2 = \|w_Q D_s f\|_2^2.$$

Hence the norm of the neglected coefficients at scale s obeys

$$\begin{aligned}\sum_{\mathcal{Q}_s'}\sum_\lambda |\alpha_{Q,\lambda}|^2 &= \sum_{\mathcal{Q}_s'} \|w_Q D_s f\|_2^2\\ &\leq C 2^{4s}\sum_{\mathcal{Q}_s'}\big(1+2^{2s}\,d(Q, Q_0)\big)^{-2m+1}\|f\|_{L^2}^2.\end{aligned}$$

Now $Q = Q(s, k_1, k_2)$ can be reparametrized in terms of k_1 and k_2, and we have $Q \in \mathcal{Q}_s'$ only if $(k_1, k_2) \notin \mathcal{K}$, where

$$\mathcal{K} = \{(k_1, k_2) : -1 \leq k_i \leq 2^s\}.$$

Define the slightly smaller index set

$$\mathcal{K}_0 = \{(k_1, k_2) : 0 \leq k_i < 2^s\}.$$

We note that for $Q \in \mathcal{Q}_s'$,

$$d(Q, Q_0) \geq c2^{-s} d\big((k_1, k_2), \mathcal{K}_0\big),$$

and that for all sufficiently large m,

$$\sum_{k \notin \mathcal{K}} d\big((k_1, k_2), \mathcal{K}_0\big)^{-2m} < \infty.$$

It follows that for such m, there is C_m with

$$2^{4s}\sum_{\mathcal{Q}'_s}\bigl(1+2^{2s}\,d(Q,Q_0)\bigr)^{-2m+1} \le C_m 2^{-s(2m-5)}.$$

We conclude

$$\sum_{\mathcal{Q}'_s}\sum_{\lambda}|\alpha_{Q,\lambda}|^2 \le C_m 2^{-s(2m-5)}\|f\|_2^2.$$

Summing now across $s \ge s_0(\varepsilon)$, we get, with $2m-5 \ge 6$,

$$\sum_{s\ge s_0}\sum_{\mathcal{Q}'_s}\sum_{\lambda}|\alpha_{Q,\lambda}|^2 \le C_m\varepsilon^{4/5}\|f\|_2^2, \tag{A.5}$$

where we used (7.1) to obtain

$$\sum_{s\ge s_0(\varepsilon)} 2^{-6s} \le C2^{-6s_0(\varepsilon)} \le C\varepsilon^{4/5}.$$

A.1.3. *Ridgelet localization.* We have just considered the collection $\mathcal{Q}'_s$ of squares Q far from Q_0. We have shown that at scales $s \ge s_0$ the combined energy in all coefficients from squares in $\mathcal{Q}'_s$ obeys the $4/5$ law. It remains to consider squares close to Q_0 and show that, while some coefficients may be large, the energy of excluded coefficients also obeys the $4/5$ law.

In the next two sections we develop a series of lemmas providing inequalities (A.8), (A.9) and (A.20) which imply the following.

COROLLARY 2. *Let Λ_s be collection of ridgelet indices $\lambda = (j,k;i,l,e)$ obeying*:

(i) $|j-(s+1)| \le 3$,
(ii) $|i-j| < s$ *and*
(iii) $|k| \le 2^{j+1}$.

For each $m \ge 0$, there exists a constant C_m such that keeping those λ's in Λ_s results in an error bounded by

$$\sum_{\lambda\notin\Lambda_s}|\alpha_{Q,\lambda}|^2 \le C_m 2^{-2sm}\|D_s f\|_2^2. \tag{A.6}$$

We now show that using this estimate and considering all squares near Q_0, we obtain the $4/5$ law. The relevant collection $\mathcal{Q}_s \setminus \mathcal{Q}'_s$ of squares Q s.t. $d(Q,Q_0) < 2^{-s+1}$ has, say, N_s squares; note that $N_s \le 2^{2s} + 2^{s+3}$. Hence from (A.6),

$$\begin{aligned}\sum_{Q\in\mathcal{Q}_s\setminus\mathcal{Q}'_s}\sum_{\lambda\notin\Lambda_s}|\alpha_{Q,\lambda}|^2 &\le \sum_{Q\in\mathcal{Q}_s\setminus\mathcal{Q}'_s} C_m 2^{-2sm}\|D_s f\|^2\\ &= C_m 2^{-2sm}\|D_s f\|^2 \sum_{Q\in\mathcal{Q}_s\setminus\mathcal{Q}'_s} 1\\ &\le C_m 2^{-2s(m-2)}\|D_s f\|^2.\end{aligned}$$

For instance, take $m = 4$ in (A.6). With this choice of m we have

$$\sum_{Q \in \mathcal{Q}_s \backslash \mathcal{Q}'_s} \sum_{\lambda \notin \Lambda_s} |\alpha_{Q,\lambda}|^2 \le C 2^{-4s} \|D_s f\|^2 \le C 2^{-6s}.$$

Therefore, since we set $2^{-6s_0} \sim \varepsilon^{4/5}$, we have

$$\sum_{s \ge s_0} \sum_{Q \in \mathcal{Q}_s \backslash \mathcal{Q}'_s} \sum_{\lambda \notin \Lambda_s} |\alpha_{Q,\lambda}|^2 \le C 2^{-6s_0} = C\varepsilon^{4/5}. \tag{A.7}$$

A.1.4. *Localization in angular scale and ridge location.* We let $a_\lambda = \langle g, \rho_\lambda \rangle$ be the ridgelet coefficient sequence of an object $g \in L_2(\mathbf{R}^2)$ which is supported in the unit square.

LEMMA 7. *Under the support constraint* $\mathrm{supp}(g) \subset [0, 1]^2$, *we have, for each* $m > 0$,

$$\sum_\lambda 2^{2(i-j)m} |a_\lambda|^2 \le C_m \|g\|_2^2 \tag{A.8}$$

and

$$\sum_{k : |k| \ge 2^{j+1}} \sum_{i,\ell,e} |a_\lambda|^2 \le C_m 2^{-2jm} \|g\|_2^2. \tag{A.9}$$

The first inequality concerns localization in angular scale i; it shows that, when the object is compactly supported, there is negligible energy at any high angular scale for any j, k. The second inequality shows that there is negligible energy at ridge locations corresponding to ridges far from the unit disk.

In fact, inequality (A.8) is proved in [8]. So we prove only (A.9) here. Our proof will be based on three lemmas. We first state the lemmas, derive their implications and only later prove the lemmas.

With $\psi_{j,k}$ the Meyer wavelet as in Section 2, define

$$\psi^+_{j,k}(t) = \frac{1}{2\pi} \int e^{i\lambda t} \hat{\psi}_{j,k}(\lambda) |\lambda| \, d\lambda. \tag{A.10}$$

Owing to the Fourier multiplier $|\lambda|$ this is a fractionally differentiated version of $\psi_{j,k}$. Using this, we can state the key *angular energy identity* as follows.

LEMMA 8.

$$\sum_{i,\ell,e} |a_\lambda|^2 = \int \left(\int Rg(t,\theta) \psi^+_{j,k}(t) \, dt \right)^2 d\theta. \tag{A.11}$$

In words, the ridgelet coefficients associated with a given j, k measure the energy of the variation in θ of the function $\theta \mapsto \langle Rg(\cdot, \theta), \psi^+_{j,k}\rangle$.

The right-hand side of identity (A.11) suggests that to get (A.9), we should show that $\langle Rg(\cdot, \theta), \psi^+_{j,k}\rangle$ is already small as soon as $|k| \geq 2 \cdot 2^j$ and that it decreases rapidly with increasing k.

Of course, the function $\psi^+_{j,k}$ is localized near the dyadic interval $[k2^{-j}, (k+1)2^{-j}]$ while the Radon transform Rg vanishes for $|t| \geq 1$ and, therefore, their interaction is indeed rather weak provided $|k| \geq 2 \cdot 2^j$ and j is large. Moreover, because of Radon isometry, $\int Rg(t, \theta)\psi^+_{j,k}(t)$ turns out to be controllable by $\|g\|_2$. The combination of these observations leads to (A.9).

The next two lemmas formalize these observations, at which point the proof becomes straightforward.

Let $(\phi_I : I = (j, k, e))$ be a nice orthonormal wavelet basis for $L^2(\mathbf{R})$, starting from coarsest level $j = j_0$, with scaling functions $\phi_{j_0,k,0}$ and wavelets $\phi_{j,k,1}$ for $j \geq j_0$. By "nice" we mean that the fine scale wavelets $\phi_{j,k,1}$ have sufficiently many vanishing derivatives and sufficiently many vanishing moments.

For an L^2 function h, let $\beta_I = \langle h, \phi_I\rangle$ denote the wavelet coefficients, and define the *Besov seminorm* of order $1/2$ by

$$\|h\|_{\dot{B}^{1/2}_{2,2}} = \left(\sum_I |\beta_I \cdot 2^{j/2}|^2\right)^{1/2}.$$

It measures roughly the energy stored in the $1/2$-order fractional derivative.

LEMMA 9.

$$\int_0^{2\pi} \|Rg(\cdot, \theta)\|^2_{\dot{B}^{1/2}_2}\, d\theta \leq C\|g\|_2^2. \tag{A.12}$$

The Besov $\dot{B}^{1/2}_{2,2}$ seminorm allows one to control the wavelet coefficients of the Radon transform, via the following.

LEMMA 10. *Let $\psi^+_{j,k}$ be the fractionally differentiated Meyer wavelet* (A.10). *Suppose that h is a function supported in* $[-1, 1]$ *and obeying* $\|h\|_2 < \infty$ *as well as* $\|h\|_{\dot{B}^{1/2}_2} < \infty$. *For each* $m > 0$ *we have a constant C_m so that*

$$|\langle \psi^+_{j,k}, h\rangle| \leq C_m \|h\|_{\dot{B}^{1/2}_{2,2}} (|k| - 2^j)_+^{-m}. \tag{A.13}$$

The last three lemmas quickly yield (A.9). Indeed, we have

$$\begin{aligned}\sum_{i,\ell,e} |a_\lambda|^2 &= \int_0^{2\pi} |\langle \psi^+_{j,k}, Rg(\cdot, \theta)\rangle|^2 d\theta \\ &\leq C_m^2 \int_0^{2\pi} \|Rg(\cdot, \theta)\|^2_{B^{1/2}_{2,2}}\, d\theta (|k| - 2^j)_+^{-2m} \\ &= C\|g\|_2^2 (|k| - 2^j)_+^{-2m}.\end{aligned}$$

For $m > 1$, we have, of course,

$$\sum_{|k|\ge 2\cdot 2^j} (|k| - 2^j)_+^{-2m} \le C_m 2^{-(2m-1)j}$$

and so

$$\sum_{|k|\ge 2\cdot 2^j} \sum_{i,\ell,e} |a_\lambda|^2 \le C_m' \|g\|_2^2 2^{-(2m-1)j},$$

completing the proof of (A.9). It remains, of course, to prove the lemmas.

PROOF OF LEMMA 8. Let $\tau_\lambda(t,\theta)$ denote the antipodally symmetrized nonorthogonal tensor wavelets $(\psi_{j,k}^+(t) w_{i,\ell}^e(\theta) + \psi_{j,k}^+(-t) w_{i,l}^e(\theta+\pi))/2$; see [22] for details. The orthoridgelet coefficients a_λ are given by analysis of the Radon transform via

$$a_\lambda = [Rg, \tau_\lambda].$$

Let now $A_{j,k}(\theta) = \langle \psi_{j,k}^+, Rg(\cdot,\theta)\rangle$. Then if $\lambda = (j,k;i,\ell,e)$,

$$\begin{aligned} a_\lambda &= \int_0^{2\pi} \big(A_{j,k}(\theta) w_{i,\ell}^e(\theta) + A_{j,1-k}(\theta) w_{i,\ell}^e(\theta+\pi)\big)/2\, d\theta \\ &= \int_0^{2\pi} \big(A_{j,k}(\theta) w_{i,\ell}^e(\theta) + A_{j,k}(\theta+\pi) w_{i,\ell}^e(\theta+\pi)\big)/2\, d\theta \qquad \text{(A.14)} \\ &= \int_0^{2\pi} A_{j,k}(\theta) w_{i,\ell}^e(\theta)\, d\theta, \end{aligned}$$

where the identities $\psi_{j,k}^+(-t) = \psi_{j,1-k}^+(t)$ and $Rg(-t, \pi+\theta) = Rg(t,\theta)$ were used. Let $\mathcal{I}(j)$ denote the set of (i, j, e) tuples obeying $i \ge j$, $0 \le \ell < 2^i$ and $e \in \{0,1\}$ if $i = j$ and $e = 1$ if $i > j$. Each collection of periodized wavelets $\{w_{i,\ell}^e : (i,\ell,e) \in \mathcal{I}(j)\}$ makes an orthonormal basis for $L^2(d\theta)$. The corresponding Parseval relation for each such basis says that for each $L^2(d\theta)$ function $A(\theta)$,

$$\sum_{\mathcal{I}(j)} |\langle A, w_{i,\ell}^e\rangle|^2 = \int_0^{2\pi} |A(\theta)|^2\, d\theta.$$

Applying this to each $A(\theta) = A_{j,k}(\theta)$,

$$\sum_{\mathcal{I}(j)} |a_\lambda|^2 = \int_0^{2\pi} A_{j,k}^2(\theta)\, d\theta.$$

Then (A.11) follows upon taking note that in (A.11), the intended range of the sum over i, ℓ, e is exactly $\mathcal{I}(j)$. □

PROOF OF LEMMA 9. In fact, more is true: the two sides are equivalent to within fixed constant multiples.

Consider the $1/2$-order L^2-Sobolev norm defined in the frequency domain by

$$\|h\|_{\dot{W}_2^{1/2}} = \int |\hat{h}(\lambda)|^2 |\lambda| \, d\lambda.$$

We have the very well-known identity,

$$\int_0^{2\pi} \|Rg(\cdot,\theta)\|^2_{\dot{W}_2^{1/2}} \, d\theta = \frac{1}{2\pi^2} \|g\|_2^2,$$

which is proved as follows:

$$\begin{aligned}\frac{1}{2}\int_0^{2\pi} \|Rg(\cdot,\theta)\|^2_{\dot{W}_2^{1/2}} \, d\theta &= \frac{1}{2}\int_0^{2\pi}\int_{-\infty}^{\infty} |\widehat{Rg}(\lambda,\theta)|^2 |\lambda| \, d\lambda \\ &= \int_0^{2\pi}\int_0^{\infty} \big|\hat{g}\big(\xi(r,\theta)\big)\big|^2 r \, dr \, d\theta \\ &= \int |\hat{g}(\xi)|^2 \, d\xi = \frac{1}{4\pi^2}\|g\|_2^2,\end{aligned}$$

where $\xi(\lambda,\theta) = (\lambda\cos(\theta), \lambda\sin(\theta))$. The second equality derives from the projection-slice theorem [40], which says that the one-dimensional Fourier transform in t of $Rg(t,\theta)$ gives the one-dimensional radial slice $\hat{g}(\xi(\lambda,\theta))$ of the Fourier transform as a function of λ.

Because one-dimensional wavelets provide a biorthogonal decomposition of the one-dimensional fractional differentiation operator, standard applications of ideas similar to those in Section 3 will show that the homogeneous Besov norm $\dot{B}^{1/2}_{2,2}$ is an equivalent norm to $\dot{W}_2^{1/2}$. This norm equivalence is of course very well known and could very well be derived by other approaches; see [52]. The result (A.12) follows. □

PROOF OF LEMMA 10. We begin by relabeling the index j, k, which stays constant throughout the proof, as j', k', allowing j, k to be used in the proof as free variables. Obviously,

$$\langle \psi^+_{j',k'}, h\rangle = \sum_I \beta_I \langle \psi^+_{j',k'}, \phi_I\rangle,$$

so we reduce matters to the study of the sequence $(\langle \psi^+_{j',k'}, \phi_I\rangle)_I$. We will establish later below the existence, for each $m > 0$, of a constant C_m so that

$$|\langle \psi^+_{j',k'}, \phi_I\rangle| \le C_m 2^{j'/2} \cdot 2^{-|j'-j|3/2}\big(1 + 2^{j'} d(t_{j',k'}, I)\big)^{-m}. \tag{A.15}$$

Assuming this and letting

$$\mathcal{I} = \{I : |\beta_I(h)| \neq 0\},$$

we have

$$
\begin{aligned}
|\langle\psi^{+}_{j',k'}, h\rangle| &= \sum_{\mathcal{I}} (\beta_I 2^{j/2})\big(\langle\psi^{+}_{j',k'}, \phi_I\rangle 2^{-j/2}\big)\\
&\le \left(\sum_{\mathcal{I}} \beta_I^2 2^j\right)^{1/2} \left(\sum_{\mathcal{I}} |\langle\psi^{+}_{j',k'}, \phi_I\rangle|^2 2^{-j}\right)^{1/2} \qquad \text{(A.16)}\\
&= \|h\|_{\dot{B}^{1/2}_{2,2}} \left(\sum_{\mathcal{I}} |\langle\psi^{+}_{j',k'}, \phi_I\rangle|^2 2^{-j}\right)^{1/2}.
\end{aligned}
$$

Let now $\bar{I}$ denote the support interval of ϕ_I. When it is necessary to specify the corresponding values of j, k, we write $\bar{I}_{j,k}$. For $j \ge j_0$ set $\mathcal{K}(j) = \{k : \bar{I}_{j,k} \cap [-1, 1] \ne \varnothing\}$. Then because supp$(h) \subset [-1, 1]$, all members of $\mathcal{I}$ with scale index j must have position index $k \in \mathcal{K}(j)$. Hence

$$
\begin{aligned}
&\sum_{\mathcal{I}} 2^{-j} |\langle\psi^{+}_{j',k'}, \phi_I\rangle|^2\\
&\qquad \le C_m^2 \sum_j \sum_{k\in\mathcal{K}(j)} 2^{-j}\Big(2^{j'} \cdot 2^{-3|j-j'|}\big(1 + 2^{j'} d(t_{j',k'}, \bar{I}_{j,k})\big)^{-2m}\Big)\\
&\qquad \le C_m^2 \sum_j 2^{-2|j-j'|} \sum_{k\in\mathcal{K}(j)} \big(1 + 2^{j'} d(t_{j',k'}, \bar{I}_{j,k})\big)^{-2m}.
\end{aligned}
$$

Now if $j \ge j'$ then from $2^{j'} d(t_{j',k'}, \bar{I}) \ge (|k'| - 2^{j'})_+$, valid for $|k| > 2 \cdot 2^{j'}$,

$$
\begin{aligned}
\sum_{k\in\mathcal{K}(j)} \big(1 + 2^{j'} d(t_{j',k'}, \bar{I}_{j,k})\big)^{-2m} &\le 2^j C \int_{-1}^{1} (1 + 2^{j'} |t_{j',k'} - u|)^{-2m}\, du\\
&\le 2^j C' \big(|k'| - 2^{j'}\big)_+^{-2m}
\end{aligned}
$$

and similarly for $j_0 \le j \le j'$, with $2^{j'}$ replacing 2^j on the extreme right-hand side. Hence,

$$
\begin{aligned}
\sum_{\mathcal{I}} 2^{-j} |\langle\psi^{+}_{j',k'}, \phi_I\rangle|^2 &\le C \sum_j 2^{-2|j-j'|} 2^{\max(j,j')} \big(|k'| - 2^{j'}\big)_+^{-2m}\\
&\le C \sum_j 2^{-|j'-j|} \big(|k'| - 2^{j'}\big)_+^{-2m} \qquad \text{(A.17)}\\
&= C'' \big(|k'| - 2^{j'}\big)_+^{-2m}.
\end{aligned}
$$

Combining (A.17) and (A.16) and relabeling $j', k' \mapsto j, k$ completes the proof of (A.13), modulo the argument for (A.15).

Returning to (A.15), we begin with the remark that, with $D = \frac{d}{dt}$ and D^n having the obvious meaning for $n = 0, \pm 1, \pm 2$, etc., we have

$$\|D^n \psi^+_{j',k'}\|_{L^\infty(\bar{I})} \le C_m\bigl(1 + 2^{j'} d(t_{j',k'}, \bar{I})\bigr)^{-m} 2^{j'n}, \qquad n = 0, \pm 1, \pm 2, \dots .$$

To obtain (A.15), assume first that $j_0 \le j \le j'$. Then, setting $n = 2$,

$$\begin{aligned} |\langle \psi^+_{j',k'}, \phi_I\rangle| &\le \|D^{-2}\psi^+_{j',k'}\|_{L^\infty(\bar{I})} \|D^2 \phi_I\|_{L^1} \\ &= C_m 2^{-j'}\bigl(1 + 2^{j'} d(t_{j',k'}, \bar{I})\bigr)^{-m} 2^{3/2j} \\ &= C_m\bigl(1 + 2^{j'} d(t_{j',k'}, \bar{I})\bigr)^{-m} 2^{j'/2} \cdot 2^{-|j-j'|3/2}. \end{aligned} \tag{A.18}$$

Assume now that $j \ge j'$. Then

$$\begin{aligned} |\langle \psi^+_{j',k'}, \phi_I\rangle| &\le \|D^{2}\psi^+_{j',k'}\|_{L^\infty(\bar{I})} \|D^{-2} \phi_I\|_{L^1} \\ &= C_m 2^{3j'}\bigl(1 + 2^{j'} d(t_{j',k'}, \bar{I})\bigr)^{-m} 2^{-5/2j} \\ &= C_m\bigl(1 + 2^{j'} d(t_{j',k'}, \bar{I})\bigr)^{-m} 2^{j'/2} \cdot 2^{-|j-j'|5/2}. \end{aligned} \tag{A.19}$$

Combining the last two displays (A.18) and (A.19) yields (A.15) and completes the proof of the lemma. □

A.1.5. *Localization in ridge scale.* We now show that distant scales can be ignored.

LEMMA 11. *For each $m > 0$, there is $C_m > 0$ so that on each fixed square Q,*

$$\sum_{\lambda : |j-(s+1)|>4} |\alpha_{Q,\lambda}|^2 \le C_m 2^{-2sm} \|D_s f\|_2^2. \tag{A.20}$$

PROOF. We renormalize the dyadic square Q to unit scale. The (Q, λ) curvelet coefficient of an object f is given by

$$\begin{aligned} \alpha_{Q,\lambda} &= \langle D_s f, \psi_{Q,\lambda}\rangle \\ &= 2^s \langle D_s f, T_Q(w\rho_\lambda)\rangle \\ &= 2^{-s}\langle w T_Q^{-1}(D_s f), \rho_\lambda\rangle. \end{aligned}$$

Define the renormalized objects $h_Q = T_Q^{-1}(D_s f)$ and $g_Q = wT_Q^{-1}(D_s f) = wh_Q$. Then

$$\alpha_{Q,\lambda} = 2^{-s}\langle g_Q, \rho_\lambda\rangle. \tag{A.21}$$

Define for general integer j the dyadic Fourier corona $\Xi_j = \{\xi : |\xi| \in [2^j, 2^{j+1}]\}$ and the *main* corona $\Xi_s^* = \{\xi : |\xi| \in [2^{s-1}, 2^{s+3}]\}$. Then, of course, Ξ_s^* combined with the dyadic coronae with $j < s - 1$ and $j > s + 2$ will cover the frequency plane.

The point of the terminology *main* corona is the following. We recall that D_s is the convolution by $\Psi_{2s} = 2^{4s}\Psi(2^{2s}\cdot)$ and $\hat{\Psi}_{2s}$ is supported on the dyadic corona $\{\xi : |\xi| \in [2^{2s-1}, 2^{2s+3}]\}$. A simple change of variables gives

$$T_Q^{-1}(D_s f) = T_Q^{-1}(\Psi_{2s} * f) = \Psi_s * T_Q^{-1} f$$

(note the change of subscript) and therefore the Fourier transform of $h_Q = T_Q^{-1}(D_s f)$ is supported on the main corona Ξ_s^*.

Even for g_Q the dominant action happens near the main corona. In fact, the Fourier transform of g_Q does decay rapidly away from the main corona Ξ_s^*. Indeed, for any $m \geq 0$, we have

$$|\hat{g}_Q|^2(\xi) \leq C2^s\bigl(1 + d(\xi, \Xi_s^*)\bigr)^{-(2m+1)}\|h_Q\|_2^2. \tag{A.22}$$

To see this, note that the relation $g_Q = wh_Q$ implies, on the Fourier side, the convolution

$$\hat{g}_Q(\xi) = \int \hat{w}(\xi - \xi')\hat{h}_Q(\xi')\, d\xi'.$$

Here $\hat{w}$ is a rapidly decaying function, that is, for each $m > 0$, $|\hat{w}(\xi)| \leq C_m(1 + |\xi|)^{-(m+1)}$. A proof of (A.22) is simply obtained by a change of variables followed by an argument similar to Lemma 6.

Now for $\xi \in \Xi_j$, and $|j - s| > 3$, $d(\xi, \Xi_s^*) \geq 2^{\max(s,j)-1}$. Hence,

$$\begin{aligned}\int_{\Xi_j} |\hat{g}_Q|^2(\xi')\, d\xi' &\leq C2^s 2^{-\max(s,j)(2m+1)}\|h_Q\|_2^2 \int_{\Xi_j} d\xi' \\ &\leq C2^s 2^{-\max(s,j)(2m+1)} 2^{2\max(s,j)}\|h_Q\|_2^2 \\ &\leq C2^{-2\max(s,j)(m-1)}\|h_Q\|_2^2.\end{aligned}$$

Finally, the elements ρ_λ of the ridgelet orthobasis are compactly supported in frequency; namely $\hat{\rho}_\lambda(\xi) = 0$ whenever $|\xi| \notin \Xi_{j-1} \cup \Xi_j$. Hence we have

$$\begin{aligned}\sum_{\lambda : |j-(s+1)|>4} |\langle g_Q, \rho_\lambda\rangle|^2 &\leq \sum_{j : |j-(s+1)|>3} \int_{\Xi_j} |\hat{g}_Q|^2(\xi')\, d\xi' \\ &\leq \sum_{j : |j-(s+1)|>3} C2^{-2\max(s,j)(m-1)}\|h_Q\|_2^2 \\ &\leq C\|h_Q\|_2^2\left(\sum_{j \leq s-3} 2^{-2s(m-1)} + \sum_{j \geq s+5} 2^{-2j(m-1)}\right) \\ &\leq C'\|h_Q\|_2^2(2^{-2s(m-2)} + 2^{-2s(m-1)}).\end{aligned}$$

To get (A.20) we now take m appropriately large. □

A.2. Sparsity of the curvelet coefficients. In this section we prove (7.4). In a separate paper [8], we derived the following upper bound on the number of coefficients whose absolute value exceeds an arbitrary cut-off $\eta > 0$:

$$\#\{\mu \in \mathcal{M}_s,\ |\alpha_\mu| \geq \eta\} \leq C \begin{cases} 0, & 2^s \geq \eta^{-2/3}, \\ \varepsilon^{-2/3}, & \eta^{-2/9} \leq 2^s \leq \eta^{-2/3}, \\ 2^{3s}, & 2^{2s} \leq \eta^{-2/9}. \end{cases}$$

A simple rescaling argument shows that

$$\#\{\mu \in \mathcal{M}_s,\ |\alpha_\mu| \geq 2^s\varepsilon\} \leq C \begin{cases} 0, & 2^s \geq \varepsilon^{-2/5}, \\ 2^{-2s/3}\varepsilon^{-2/3}, & \varepsilon^{-2/11} \leq 2^s \leq \varepsilon^{-2/5}, \\ 2^{3s}, & 2^{2s} \leq \varepsilon^{-2/11}. \end{cases}$$

From the last inequality, one easily deduces that

$$\sum_{\mu \in \mathcal{M}_s} \min(|\alpha_\mu|^2, 2^{2s}\varepsilon^2) \leq 2^{4s/3}\varepsilon^{4/3}.$$

Hence,

$$\begin{aligned} \sum_{\mu \in \mathcal{N}(\varepsilon)} \min(|\alpha_\mu|^2, 2^{2s}\varepsilon^2) &\leq \sum_{s:\, 2^s \leq \varepsilon^{-2/5}} \sum_{\mu \in \mathcal{M}_s} \min(|\alpha_\mu|^2, 2^{2s}\varepsilon^2) \\ &\leq C \sum_{s:\, 2^s \leq \varepsilon^{-2/5}} 2^{4s/3}\varepsilon^{4/3} \\ &\leq C\varepsilon^{4/5}, \end{aligned}$$

which is what needed to be shown.

A.3. Cardinality of $\mathcal{N}(\varepsilon)$. We now establish (7.5). For a scale $s \geq s_0$, the number of coefficients that one keeps per dyadic square Q equals the cardinality of Λ_s which is bounded by $C2^{3s}$. The subset $\mathcal{N}(\varepsilon)$ counts a maximum of $O(2^{2s})$ of squares Q at such scale and, therefore,

$$\#\{\mu \in \mathcal{N}(\varepsilon) \text{ s.t. } Q \in \mathcal{Q}_s\} \leq C2^{5s}.$$

Then, of course,

$$\#\mathcal{N}(\varepsilon) = \sum_{s:\, 2^s \leq \varepsilon^{-2/5}} \#\{\mu \in \mathcal{N}(\varepsilon) \text{ s.t. } Q \in \mathcal{Q}_s\} \leq C\varepsilon^{-2},$$

which proves (7.5).

A.4. Remarks. The above estimates complete the proof of Theorem 5.

Because the parameter m may be chosen arbitrarily large in (A.8), it is not difficult to show that the cardinality can be bounded by $C_\delta \varepsilon^{-8/5+\delta}$, for any $\delta > 0$.

Finally, the argument may be adapted to other choices of s_0. The description of $\mathcal{N}(\varepsilon)$ would involve a different and somewhat more complicated selection of parameters for small values of the scale s, $s_0 \le s \le s_\varepsilon$.

REFERENCES

[1] BERTERO, M. (1989). Linear inverse and ill-posed problems. In *Advances in Electronics and Electron Physics* (P. W. Hawkes, ed.). Academic Press, New York.

[2] BERTERO, M., DE MOL, C. and PIKE, E. R. (1985). Linear inverse problems with discrete data I: General formulation and singular system analysis. *Inverse Problems* **1** 301–330.

[3] BROWN, L. D. and LOW, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398.

[4] CANDÈS, E. J. (1998). Ridgelets: theory and applications. Ph.D. dissertation, Dept. Statistics, Stanford Univ.

[5] CANDÈS, E. J. (1999). Harmonic analysis of neural networks. *Appl. Comput. Harmon. Anal.* **6** 197–218.

[6] CANDÈS, E. J. (1999). Monoscale ridgelets for image compression and denoising. Unpublished manuscript.

[7] CANDÈS, E. J. and DONOHO, D. L. (1999). Ridgelets: a key to high-dimensional intermittency? *Philos. Trans. Roy. Soc. London Ser. A* **357** 2495–2509.

[8] CANDÈS, E. J. and DONOHO, D. L. (1999). Curvelets. Unpublished manuscript. Available at www.stat.stanford.edu/˜donoho/Reports/1999/curvelets.pdf

[9] CANDÈS, E. J. and DONOHO, D. L. (2000). Curvelets: A surpisingly effective nonadaptive representation of objects with edges. In *Curve and Surface Fitting* (A. Cohen, C. Rabut and L. L. Schumaker, eds.) 105–120. Vanderbilt Univ. Press, Nashville, TN.

[10] CHARBONNIER, P., BLANC-FÈRAUD, L., AUBERT, G. and BARLAUD, M. (1997). Deterministic edge-preserving regularization in computed imaging. *IEEE Trans. Image Process.* **6** 298–311.

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

[12] DAVISON, M. E. (1981). A singular value decomposition for the Radon transform in n-dimensional space. *Numer. Funct. Anal. Optim.* **3** 321–340.

[13] DEANS, S. R. (1983). *The Radon Transform and Some of Its Applications*. Wiley, New York.

[14] DOBSON, D. and SANTOSA, F. (1996). Recovery of blocky images from noisy and blurred data. *SIAM J. Appl. Math.* **56** 1181–1198.

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

[16] DONOHO, D. L. (1994). Statistical estimation and optimal recovery. *Ann. Statist.* **22** 238–270.

[17] DONOHO, D. L. (1994). Asymptotic minimax risk for sup norm loss: Solution via optimal recovery. *Probab. Theory Related Fields* **99** 145–170.

[18] DONOHO, D. L. (1995). Nonlinear solution of linear inverse problems by wavelet–vaguelette decomposition. *Appl. Comput. Harmon. Anal.* **2** 101–126.

[19] DONOHO, D. L. (1995). Denoising by soft-thresholding. *IEEE Trans. Inform. Theory* **41** 613–627.

[20] DONOHO, D. L. (1997). Renormalizing experiments for nonlinear functionals. In *Festschrift for Lucien Le Cam* (D. L. Pollard, E. N. Torgersen and G. L. Yang, eds.) 167–181. Springer, New York.

[21] DONOHO, D. L. (1999). Wedgelets: nearly minimax estimation of edges. *Ann. Statist.* **27** 859–897.

[22] DONOHO, D. L. (2000). Orthonormal ridgelets and linear singularities. *SIAM J. Math. Anal.* **31** 1062–1099.

[23] DONOHO, D. L. (2001). Sparse components of images and optimal atomic decompositions. *Constr. Approx.* **17** 353–382.

[24] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Minimax risk over ℓ_p balls for ℓ_q losses. *Probab. Theory Related Fields* **99** 277–303.

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

[26] DONOHO, D. L. and JOHNSTONE, I. M. (1995). Empirical atomic decomposition. Unpublished manuscript.

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

[28] DONOHO, D. L. and JOHNSTONE, I. M. (1999). Asymptotic minimaxity of wavelet estimators with sampled data. *Statist. Sinica* **9** 1–32.

[29] DONOHO, D. L. and LIU, R. C. (1991). Geometrizing rates of convergence III. *Ann. Statist.* **19** 668–701.

[30] DONOHO, D. L. and LOW, M. G. (1992). Renormalization exponents and optimal pointwise rates of convergence. *Ann. Statist.* **20** 944–970.

[31] DONOHO, D. L. and NUSSBAUM, M. (1990). Minimax quadratic estimation of a quadratic functional. *J. Complexity* **6** 290–323.

[32] EFROIMOVICH, S. YU. and PINSKER, M. S. (1981). Estimation of square-integrable density on the basis of a sequence of observations. *Problemy Peredachi Informatsii* **17** 50–68 (in Russian). *Problems Inform. Transmission* **17** (1982) 182–196 (in English).

[33] EFROIMOVICH, S. YU. and PINSKER, M. S. (1982). Estimation of square-integrable probability density of a random variable. *Problemy Peredachi Informatsii* **18** 19–38 (in Russian). *Problems Inform. Transmission* **18** (1983) 175–189 (in English).

[34] EFROMOVICH, S. and SAMAROV, A. (1996). Asymptotic equivalence of nonparametric regression and white noise models has its limits. *Statist. Probab. Lett.* **28** 143–145.

[35] FRAZIER, M. and JAWERTH, B. (1985). Decomposition of Besov spaces. *Indiana Univ. Math. J.* **34** 777–799.

[36] FRAZIER, M. and JAWERTH, B. (1990). A discrete transform and decompositions of distribution spaces. *J. Funct. Anal.* **93** 34–170.

[37] FRAZIER, M., JAWERTH, B. and WEISS, G. (1991). *Littlewood–Paley Theory and the Study of Function Spaces*. Amer. Math. Soc., Providence, RI.

[38] GEMAN, D. and REYNOLDS, G. (1992). Constrained restoration and the recovery of discontinuities. *IEEE Trans. Pattern Anal. Mach. Intell.* **14** 367–382.

[39] GEMAN, D. and YANG, C. (1995). Nonlinear image recovery with half-quadratic regularization. *IEEE Trans. Image Processing* **4** 932–946.

[40] HELGASON, S. (1980). *The Radon Transform*. Birkhäuser, Boston.

[41] IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1984). Nonparametric estimation of the value of a linear functional in a Gaussian white noise. *Theory Probab. Appl.* **29** 18–32.

[42] JOHNSTONE, I. M. and SILVERMAN, B. W. (1990). Speed of estimation in positron emission tomography and related inverse problems. *Ann. Statist.* **18** 251–280.

[43] JOHNSTONE, I. M. and SILVERMAN, B. W. (1991). Discretization effects in statistical inverse problems. *J. Complexity* **7** 1–34.

[44] JOHNSTONE, I. M. and SILVERMAN, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *J. Roy. Statist. Soc. Ser. B.* **59** 319–351.

[45] JONSSON, E., HUANG, S. C. and CHAN, T. (1998). Total variation regularization in positron emission tomography. Technical report, Dept. Mathematics, Univ. California, Los Angeles.

[46] KALIFA, J. and MALLAT, S. G. (1999). Thresholding estimators for inverse problems and deconvolutions. Unpublished manuscript.

[47] 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.

[48] KOROSTELEV, A. P. and TSYBAKOV, A. B. (1993). *Minimax Theory of Image Reconstruction*. *Lecture Notes in Statist.* **82**. Springer, New York.

[49] LEMARIÉ, P. G. and MEYER, Y. (1986). Ondelettes et bases hilbertiennes. *Rev. Mat. Iberoamericana* **2** 1–18.

[50] LOUIS, A. K. (1986). Incomplete data problems in X-ray computerized tomography I. Singular value decomposition of the limited-angle transform. *Numer. Math.* **48** 251–262.

[51] MALLAT, S. G. (1989). Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$. *Trans. Amer. Math. Soc.* **315** 69–87.

[52] MEYER, Y. (1990). *Ondelettes et Opérateurs I. Ondelettes*. Hermann, Paris.

[53] MEYER, Y. (1990). *Ondelettes et Opérateurs II. Opérateurs de Calderón Zygmund*. Hermann, Paris.

[54] NUSSBAUM, M. (1985). Spline smoothing in regression models and asymptotic efficiency in L_2. *Ann. Statist.* **13** 984–997.

[55] O'SULLIVAN, F. (1986). A statistical perspective on ill-posed inverse problems. *Statist. Sci.* **1** 502–527.

[56] PILZ, J. (1986). Minimax linear regression estimation with symmetric parameter restrictions. *J. Statist. Plann. Inference* **13** 297–318.

[57] PINSKER, M. S. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problemy Peredachi Informatsii* **16** 52–68 (in Russian). *Problems Inform. Transmission* **16** (1980) 120–133 (in English).

[58] RUDIN, L., OSHER, S. and FATEMI, E. (1992). Nonlinear total variation based noise removal algorithms. *Phys. D* **60** 259–268.

[59] STEIN, E. M. (1970). *Singular Integrals and Differentiability Properties of Functions*. Princeton Univ. Press.

[60] TEBOUL, S., BLANC-FÉRAUD, L., AUBERT, G. and BARLAUD, M. (1998). Variational approach for edge-preserving regularization using coupled PDEs. *IEEE Trans. Image Processing* **7** 387–397.

[61] TERZOPOULOS, D. (1986). Regularization of inverse visual problems involving discontinuities. *IEEE Trans. Pattern Anal. Machine Intell.* **8** 413–424.

[62] TIKHONOV, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Doklady* **4** 1035–1039.

[63] VETTERLI, M. and KOVACEVIC, J. (1995). *Wavelets and Subband Coding*. Prentice-Hall, Englewood Cliffs, NJ.

[64] WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

DEPARTMENT OF APPLIED MATHEMATICS
CALIFORNIA INSTITUTE OF TECHNOLOGY
MAIL CODE 217-50
PASADENA, CALIFORNIA 91125