Journal of Approximation Theory **111**, 143–179 (2001)
doi:10.1006/jath.2001.3568, available online at http://www.idealibrary.com on IDEAL®

Ridge Functions and Orthonormal Ridgelets

David L. Donoho

Department of Statistics, Stanford University, Stanford, California 94305-4065

Communicated by Will Light

Received February 22, 1999; accepted in revised form February 20, 2001;
published online June 18, 2001

Orthonormal ridgelets provide an orthonormal basis for $L^2(\mathbf{R}^2)$ built from special angularly-integrated ridge functions. In this paper we explore the relationship between orthonormal ridgelets and true ridge functions $r(x_1 \cos\theta + x_2 \sin\theta)$. We derive a formula for the ridgelet coefficients of a ridge function in terms of the 1-D wavelet coefficients of the ridge profile $r(t)$. The formula shows that the ridgelet coefficients of a ridge function are heavily concentrated in ridge parameter space near the underlying scale, direction, and location of the ridge function. It also shows that the rearranged weighted ridgelet coefficients of a ridge function decay at essentially the same rate as the rearranged weighted 1-D wavelet coefficients of the 1-D ridge profile $r(t)$. In short, the full ridgelet expansion of a ridge function is in a certain sense equally as sparse as the 1-D wavelet expansion of the ridge profile.

It follows that partial ridgelet expansions can give good approximations to objects which are countable superpositions of well-behaved ridge functions. We study the nonlinear approximation operator which "kills" coefficients below certain thresholds (depending on angular- and ridge-scale); we show that for approximating objects which are countable superpositions of ridge functions with 1-D ridge profiles in the Besov space $B^{1/p}_{p,p}(\mathbf{R})$, $0 < p < 1$, the thresholded ridgelet approximation achieves optimal rates of N-term approximation. This implies that appropriate thresholding in the ridgelet basis is equally as good, for certain purposes, as an ideally-adapted N-term nonlinear ridge approximation, based on perfect choice of N-directions.

Key Words: wavelets; ridge function; ridgelet; radon transform; best N-term approximation; thresholding of wavelet coefficients.

1. INTRODUCTION

In [9], we introduced orthonormal ridgelets, defined as follows. Let $(\psi_{j,k}(t): j \in \mathbf{Z}, k \in \mathbf{Z})$ be an orthonormal basis of Meyer wavelets for $L^2(\mathbf{R})$ [14], [15, Engl. Transl. p. 75], [6, pp. 66–70], and let $(w^0_{i_0,\ell}(\theta), \ell = 0, ..., 2^{i_0} - 1; w^1_{i,\ell}(\theta), i \geqslant i_0, \ell = 0, ..., 2^i - 1)$ be an orthonormal basis for $K^2[0, 2\pi)$ made of periodized Lemarié scaling functions $w^0_{i_0,\ell}$ at level i_0 and periodized Meyer wavelets $w^1_{i\ell}$ at levels $i \geqslant i_0$ [15, Engl. Transl. p. 113]. (We suppose a particular normalization of these functions; see Section 2.1

0021-9045/01 $35.00

below). Let $\hat{\psi}_{j,k}(\omega)$ denote the Fourier transform of $\psi_{j,k}(t)$, and define ridgelets $\rho_\lambda(x)$, $\lambda=(j, k; i, \ell, \varepsilon)$ as functions of $x\in\mathbf{R}^2$ using the frequency-domain definition

$$\hat{\rho}_\lambda(\xi)=|\xi|^{-1/2}\,(\hat{\psi}_{j,k}(|\xi|)\,w^{\varepsilon}_{i,\ell}(\theta)+\hat{\psi}_{j,k}(-|\xi|)\,w^{\varepsilon}_{i,\ell}(\theta+\pi))/2. \qquad (1.1)$$

Here the indices run as follows: $j,\ k\in\mathbf{Z},\ \ell=0, ..., 2^{i-1}-1,\ i\geqslant i_0$; and, if $i>i_0$, $i\geqslant j$. Also, if $i>i_0$ and $i>j$, then necessarily $\varepsilon=1$. Let Λ denote the set of all such indices λ. (To avoid confusion, note that in [9] two orthonormal systems were discussed; the one we study here was defined and used throughout the main body of [9]). In that article, it was shown that this collection of functions makes an orthonormal set for $L^2(\mathbf{R}^2)$.

Define now $\psi^+_{j,k}(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}|\omega|^{1/2}\,\hat{\psi}_{j,k}(\omega)\,e^{i\omega t}\,d\omega$; this is a fractionally-differentiated Meyer wavelet. The cited article also showed that for $x=(x_1, x_2)\in\mathbf{R}^2$,

$$\rho_\lambda(x)=\frac{1}{4\pi}\int_0^{2\pi}\psi^+_{j,k}(x_1\cos\theta+x_2\sin\theta)\,w^{\varepsilon}_{i,\ell}(\theta)\,d\theta. \qquad (1.2)$$

Here, each $\psi^+_{j,k}(x_1\cos\theta+x_2\sin\theta)$ is a *ridge* function of $x\in\mathbf{R}^2$, i.e., a function of the form $r(x_1\cos\theta+x_2\sin\theta)$. Therefore ρ_λ is obtained by "averaging" ridge functions with ridge angles θ localized near $\theta_{i,\ell}=2\pi\ell/2^i$; this justifies the "ridgelet" appellation.

1.1. *Ridge Functions*

The ridge function terminology was introduced in the 1970's by Logan and Shepp [13] in connection with the mathematics of computed tomography. In recent years, ridge functions have appeared often in the literature of approximation theory and statistics as part of methodological topics such as neural networks [1] and projection pursuit regression [11]. A celebrated result of Barron [1], building on work of Lee Jones [12], established the ability of an appropriately-chosen superposition of ridge functions $\hat{f}_N(x)=\sum_{n=1}^{N}a_n\sigma(u'_nx-b_n)$ to converge at good rates to an underlying function f obeying a certain smoothness condition, even in high dimensions [1]. A variety of interesting literature has ensued.

Candès [3] has pointed out a key drawback of much of the work on neural nets: the lack of a constructive, stable character. Many results are of the form "there exists a sequence of approximations"; they don't exhibit a sequence of approximations concretely. Other results, even if constructive, fail to exhibit stability. For example the directions u_n, locations b_n and coefficients a_n appearing in an N-term approximation are not stable functionals of the underlying approximant f: small changes in f can lead to large

changes in these functionals, which are therefore not interpretable as having any reliable meaning.

In an instructive analogy, the situation in much of the literature on ridge function approximation could be compared to a (non-existent) situation where we knew that there exist approximations to functions by superpositions of plane waves $\exp\{\sqrt{-1}\,\omega' x\}$, but where Fourier analysis had not been invented, so we didn't know how to construct stable approximations, or, more properly, didn't know that such stable approximations were possible in principle. Equally, we could imagine a situation where we knew that there exist approximation by superpositions of wavelets $\psi_{a,b}(x) = a^{1/2}\psi(ax-b)$, but where orthonormal wavelet analysis had not been invented, and so we didn't know how to construct stable approximations.

1.2. *Ridgelets*

Candès [3] and Murata [16] independently obtained results suggesting that, just as the Fourier transform allows to construct stable approximations by plane waves, and the wavelet transform allows to construct stable approximations by wavelets, there is a transform which allows to give stable approximations by special ridge functions. They developed the *continuous ridgelet transform*, defined using ridge-wavelets $\psi_{a,b,u}(x) = \psi(au'x-b)\,a^{1/2}$ where $a>0$, $b\in R$, and $\in S^{d-1}$, and taking values $\mathscr{R}_f(a,b,u) = \langle f, \psi_{a,b,u}\rangle$ (we are following notation of [3]). This continuous transform has a reconstruction formula $f(x) = \int \mathscr{R}_f(a,b,\theta)\,\psi_{a,b,\theta}(x)\,d\mu(a,b,\theta)$, representing f as a continuous superposition of ridge-wavelets (here $d\mu$ is an appropriate reference measure). Candès showed that the coefficients of this continuous transform were stable—obeying a Parseval relation. He also showed that discrete decompositions were possible, so that for L^2 spaces of compactly supported functions one could develop a *frame of ridgelets*—a discrete family $(\psi_{a_n,b_n,u_n}(x))$ serving the role of an approximating system. The outcome of this pioneering research was, more or less, to show that ridgelet analysis could be used for nonlinear approximation in a way paralleling Fourier analysis and wavelet analysis; one could construct good approximations from superpositions of N ridge functions by a very simple algorithm: simply form a partial reconstruction taking the N most significant coefficients in the ridgelet frame. However, in Candès' approach, certain interpretational details remained to be clarified. Because the ridgelet frames were not tight, the dual frame elements form part of the complete picture; but as these were known only implicitly, certain properties of the algorithm remain only partially accessible to analysis.

1.3. *Orthonormal Ridgelets*

The "classic ridgelets" of Candès are not in $L^2(\mathbf{R}^2)$, being constant on lines $t = x_1\cos\theta + x_2\sin\theta$ in the plane. This fact seems responsible for

certain technical difficulties in the deployment and interpretation of discrete systems based on Candès notion of ridgelet. In [9] the author had the subversive idea to broaden the concept of ridgelet somewhat, allowing "wide-sense" ridgelets to be functions obeying certain localization properties in a radial frequency × angular frequency domain. Under this broader conception, ridgelets no longer are of the form $\psi_{a,b,u}(x)$, so the elegant simplicity of formulation is lost. However, in exchange, it becomes possible to have an orthonormal set of "wide-sense" ridgelets. These "orthonormal ridgelets" are believed to be appropriate L^2-substitutes for ridge functions, and to fulfill the goal of a constructive and stable system which *although not based on true ridge functions* are believed to play operationally the same role as ridge functions.

1.4. *A Fruitful Analogy*

A certain analogy may help the reader understand the situation. Suppose we wanted to approximate a function $f(t)$ of a single real variable by a finite superposition $\tilde{f}_n(t) = \sum_{n=1}^{N} c_n \phi_{a_n, b_n}(t)$ of Gaussian bumps $\phi_{a,b}(t) = \phi(at - b)$. How could we go about this? As Pencho Petrushev has pointed out to the author, there is no known method for constructively obtaining a stable and effective N-term approximation by such superpositions of Gaussians.

On the other hand, if we switch attention to orthonormal wavelets, which are not Gaussians, we obtain a discrete scale-location family where good N-term approximations are easy to construct; one simply uses the N biggest terms in the wavelet expansion. Moreover, one ultimately gains in the practical domain as well, because fast wavelet transform algorithms become available. The N-term expansion is no longer an expansion in Gaussians; it has been replaced by what are in a sense quasi-Gaussians, but for which the theoretical and practical questions are much better posed and are solvable.

In facing Petrushev's question, we therefore make decisive progress by abandoning insistence on the original specification of the synthesizing elements, and passing to a new system of "similar" elements, where the answers are clear and clean.

This is our philosophy in the present article; we explore the idea that one can understand ridge function approximation not by studying approximation through narrow-sense ridgelets (as in Candès' papers [3, 4]) but by abandoning these, and studying approximation through a wide-sense ridgelet system—the orthonormal ridgelets.

1.5. *A Special Relationship*

There is in fact a close connection between orthonormal ridgelets and certain special ridge functions.

Let $\gamma = (j_0, k_0, \theta_0)$ be given, and define the *special ridge function* $r^\gamma(x) = \psi^+_{j_0, k_0}(x_1 \cos\theta_0 + x_2 \sin\theta_0)$. Define, for $\lambda = (j, k; i, \ell, \varepsilon)$, the array of coefficients

$$a^\gamma_\lambda = \delta_{j_0 j}\delta_{k_0 k} \cdot w^e_{i,\ell}(\theta_0) + \delta_{j_0 j}\delta_{k_0, 1-k} \cdot w^e_{i,\ell}(\theta_0 + \pi), \tag{1.3}$$

where the δ's denote Kronecker symbols. We will show below that the special ridge function r^γ has a representation by ridgelets ρ_λ using precisely the coefficients a^γ_λ:

$$r^\gamma = \sum_\lambda a^\gamma_\lambda \rho_\lambda. \tag{1.4}$$

The special form of the coefficient array $(a^\gamma_\lambda : \lambda \in \Lambda)$ is remarkable: it is very sparse. In the j and k indices all the action occurs at precisely $j = j_0$ and $k = k_0$ or $k = 1 - k_0$, while in the ℓ index, all the action occurs "near" the index ℓ_0 with $\theta_{i,\ell}$ closest to θ_0, or at its counterpart closest to $\theta_0 + \pi$. In a certain sense, the pure ridge function r^γ is a superposition of a small number of ridgelets.

An apparent exception to our sparsity assertion occurs in the index i, where the coefficients a^γ_λ are increasing exponentially in i (as $w^e_{i,\ell_0}(\theta_0) \approx 2^{i/2}$). Closer inspection reveals that effective support is sparse here as well. The key is that the angular resolution index i associated with the ridgelet parameter $\lambda = (j, k; i, \ell, \varepsilon)$ measures in effect the distance of the support of ρ_λ from the origin. Hence, if we restrict attention to a disk $D = \{(x_1, x_2) : x_1^2 + x_2^2 < d^2\}$ then, in a certain sense most of the action in ρ_λ will be occurring outside the disk for d fixed and i large. Consequently, from the point of view of the contribution of the product $a^\gamma_\lambda \cdot \rho_\lambda$ within the disk D, most of the action will be concentrated at a small range in the i index as well—near $i = j_0$ when j_0 is large.

1.6. *Ridgelet Coefficients of a Ridge Function*

The connection indicated by (1.3)–(1.4) leads to a number of instructive results about the connection between orthonormal ridgelets and ridge functions. Starting now, "ridgelet" will always mean one of the orthonormal ridgelets as defined in (1.1). Our initial result will be a formula for the ridgelet coefficients of a ridge function with sufficiently-regular profile.

The notion of regularity of a profile we use in this paper is based on *homogeneous Besov spaces*. We say that a distribution r belongs to $\dot{B}^s_{p,p}(\mathbf{R})$ if it can be represented as a wavelet series

$$r = \sum_{j,k} \alpha_{j,k} \psi_{j,k}, \tag{1.5}$$

where $(\psi_{j,k})$ is a family of Meyer orthonormal wavelets, and where the coefficients $(\alpha_{j,k})$ obey

$$\left(\sum_{j,k} |\alpha_{j,k}|^p \, 2^{j(s+1/2-1/p)\,p}\right)^{1/p} < \infty. \tag{1.6}$$

We will only be interested in the cases where $p \leqslant 1$ and $s \geqslant 1/p$. It is easy to see that in such cases, (1.6) implies that (1.5) is absolutely summable to a uniformly continuous function.

• *Remark for nonspecialists.* The most well-known example of this type of space is the Bump Algebra (Meyer, 1990). This is the case $s = p = 1$ of the above family. The Bump Algebra can also be described as follows. It is the class of all functions r representable as a superposition of Gaussian bumps $g_{a,b}(t) = g(a(t-b))$, with $g(t) = e^{-t^2}$, where $r = \sum_i c_i g_{a_i, b_i}(t)$ and $\sum_i |c_i| < \infty$. This is a superposition of bumps of all possible widths and locations, where the total sum of heights of all the bumps is finite. It will turn out for some of our results that this space plays the role of a critical case, the least-regular space where our approach works.

• *Remark for specialists.* This notion of homogeneous Besov space is admittedly nonstandard, since in general such spaces consist of equivalence classes of distributions rather than classes of proper functions. In effect, our approach singles out one specific member of an equivalence class; in fact the one vanishing at $\pm\infty$. To remind ourselves of this fact, we will typically say that an r belongs to $\dot{B}^s_{p,p}$ and vanishes at $\pm\infty$.

Below we will write $\|r\|_{\dot{B}^s_{p,p}(\mathbf{R})}$ for the norm of the profile; we mean by this precisely the left side of (1.6).

Define the fractionally-integrated Meyer wavelet

$$\psi^-_{j,k}(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\omega|^{-1/2} \, \hat{\psi}_{j,k}(\omega) \, e^{i\omega t} \, d\omega.$$

THEOREM 1.1. *Let $r(t)$ be a function of a single variable, vanishing at $t = \pm\infty$ and belonging to the Bump Algebra (Besov space $\dot{B}^1_{1,1}$). Let $r_\theta(x) = r(x_1 \cos\theta + x_2 \sin\theta)$ denote the corresponding ridge function with ridge profile r and direction parameter θ. Then*

$$\langle r_\theta, \rho_\lambda \rangle = (\langle r, \psi^-_{j,k} \rangle \, w^e_{i,\ell}(\theta) + \langle r, \psi^-_{j,1-k} \rangle \, w^e_{i,\ell}(\theta+\pi))/2. \tag{1.7}$$

Here the scalar product on the left of (1.7),

$$\langle r_\theta, \rho_\lambda \rangle \equiv \int_{\mathbf{R}^2} r_\theta(x) \, \rho_\lambda(x) \, dx,$$

may be interpreted as a pairing between $L^\infty(\mathbf{R}^2)$ and $L^1(\mathbf{R}^2)$, while the scalar product on the right, $\langle r, \bar{\psi}_{j,k} \rangle = \int_{\mathbf{R}} r(t)\, \bar{\psi}_{j,k}(t)\, dt$, is a pairing between $L^\infty(\mathbf{R})$ and $L^1(\mathbf{R})$. The formula (1.7) shows that the ridgelet coefficients of a ridge function separate into two factors: (1-d wavelet coefficients of the ridge profile) $\times$ (wavelet point evaluations at the given ridge direction).

1.7. *Linear Approximation by Ridgelets*

We now consider linear approximation of ridge functions by index-limited ridgelet expansions. For a bounded function $f(x)$, let $A^{i_1} f = \sum_{i,\, j \leqslant i_1} \langle f, \rho_\lambda \rangle\, \rho_\lambda$ be the partial sum approximation to f by ridgelets with both i and j index-limited so that only scales larger than 2^{-i_1} are used.

THEOREM 1.2. *If the ridge profile $r \in \dot{B}^1_{1,1}(\mathbf{R})$ and r vanishes at $\pm\infty$, then the approximation $A^{i_1} r_\theta$ to the ridge function r_θ is well-defined and converges to r in $L^\infty(D)$-norm as $i_1 \to \infty$. If, in addition, the profile $r \in \dot{B}^s_{1,1}$, $s > 1$, then*

$$\|r_\theta - A^{i_1} r_\theta\|_{L^\infty(D)} \leqslant C 2^{-i_1(s-1)}\, \|r\|_{\dot{B}^s_{1,1}(\mathbf{R})}, \qquad i_1 = i_0, i_0 + 1, \ldots. \quad (1.8)$$

Here $L^\infty(D)$ means supremum norm over a fixed disk, $D = \{x_1^2 + x_2^2 < d^2\}$.

This shows that linear approximation by ridgelets converges rapidly to nice ridge functions. We note that the rate is comparable to what one would expect for a general method of approximation to a general function of the same smoothness.

1.8. *Nonlinear Approximation by Ridgelets*

We now consider nonlinear approximation to ridge functions by N-term ridgelet approximations. For this result, let $\eta_\delta(y, s) = y 1_{\{y \cdot s > \delta\}}$ be a thresholding function with a "scaling" argument s, allowing for adjustment of the threshold. For a bounded function f, with a finite value for

$$\bar{N}(\delta) = \sum_{\Lambda} 1_{\{|\langle f, \rho_\lambda \rangle| \cdot \|\rho_\lambda\|_{L^\infty(D)} > \delta\}}, \quad (1.9)$$

set

$$\tilde{f}_\delta = \sum_{\Lambda} \eta_\delta(\langle f, \rho_\lambda \rangle, \|\rho_\lambda\|_{L^\infty(D)})\, \rho_\lambda. \quad (1.10)$$

THEOREM 1.3. *Let the ridge profile $r(t)$ belong to Besov space $\dot{B}^s_{p,p}(\mathbf{R})$, with $s = 1/p$ and $0 < p < 1$, and vanish at $\pm\infty$. Let $r_\theta(x)$ denote the corresponding ridge function of $x \in \mathbf{R}^2$. Define a sequence of approximants $(\bar{r}_N)$ by*

letting $\bar{r}_{\bar{N}(\delta)}$ *be the* $\bar{N}(\delta)$*-term ridgelet approximation obtained from* (1.10), *with* $f = r_\theta$. *Then*

$$\|r_\theta - \bar{r}_N\|_{L^\infty(D)} \leqslant C \cdot \|r\|_{\dot{B}^s_{pp}(\mathbf{R})} \cdot N^{-(s-1)}, \qquad N = 1, 2, \ldots. \tag{1.11}$$

This shows that ridgelet thresholding is a natural procedure to construct finite sums of ridgelets converging rapidly to smooth ridge functions.

The apparently nonstandard spaces $\dot{B}^s_{p,p}$ with $p < 1$ are in fact very natural for this result; it is well-understood that these are the appropriate spaces for understanding the properties of nonlinear approximation using the L^∞ norm in $\mathbf{R}$ [7, 17]. As described below, nice enough ridge profiles r have convergent expansions $r = \sum c_{j,k} \psi^+_{j,k}$. Suppose we constructed N-term 1-d wavelet approximations to the ridge profile function r which were of the form $r_N = \sum_{n=1}^N c_{j_n,k_n} \psi^+_{j_n,k_n}$ by the device of picking the N-most "important" terms according to the value of the product $|c_{j,k}| \cdot \|\psi^+_{j,k}\|_{L^\infty[-d,d]}$. Then the hypothesis $r \in B^s_{p,p}[-d, d]$ alone would allow at best a conclusion of the form

$$\|r - r_N\|_{L^\infty[-d,d]} \leqslant \text{Const} \cdot N^{-(s-1)}, \qquad N = 1, 2, \ldots. \tag{1.12}$$

Moreover, no other stable scheme of N-term approximation (e.g. one based on something other than the wavelets $(\psi^+_{j,k})$, for example rational approximation, free knot spline approximation, etc.) can do essentially better than this—this is the meaning of the statement that such methods all have $B^s_{p,p}[-d, d]$ for their approximation spaces [7].

In this light, Theorem 1.3 is rather interesting. It says that ridgelet thresholding—which does not "know" the direction of the ridge—does essentially as well as a kind of ideally direction-adapted ridge approximation which "knows θ", Specifically, consider an N-term ridge approximation making use of θ as follows. With r_N an N-term approximation to the ridge profile, just approximate r_θ using the ridge function $r_{N,\theta}$ in exactly the right direction: $r_{N,\theta}(x) = r_N(x_1 \cos\theta + x_2 \sin\theta)$.

The properties of this ideally-adapted approximation are easily derived. When approximating ridge functions by ridge functions, the L^∞-error of 2-dimensional approximation is isometric to the L^∞-error of 1-dimensional approximation:

$$\|r - r_N\|_{L^\infty[-d,d]} = \|r_\theta - r_{N,\theta}\|_{L^\infty(D)}. \tag{1.13}$$

Combining (1.12) with (1.13), we see that the ideally-adapted N-term scheme gives

$$\|r_\theta - r_{n,\theta}\|_{L^\infty(D)} \leqslant \text{Const} \cdot N^{-(s-1)}, \qquad N = 1, 2, \ldots, \tag{1.14}$$

and this is the best one can hope for in general. Indeed, recalling the discussion after (1.12), no other way of generating N-term approximate ridge profiles r_N and then using these to build ridge functions $r_{N,\theta}$ is going to admit of substantially better estimates, uniformly over the class of all profiles r obeying $\|r\|_{\dot{B}^{1/p}_{p,p}(\mathbf{R})} \leq R$.

Comparing (1.14) with (1.11) shows that, in a certain sense, an ideal approximation to a ridge function r_θ using N pure ridge functions in the exact ridge direction does not do better than our N-term ridgelet approximation, which has no "knowledge" of θ and does not depend in any way on an assumption that the approximant is a ridge function.

1.9. *Superpositions of Ridge Functions*

Our results have immediate implications for approximating more general functions f.

Consider the linear approximation result, Theorem 1.2. Convexity of the $L^\infty(\mathbf{R}^2)$ norm, combined with linearity of A^{i_1}, immediately yields the following

COROLLARY 1.4. *Suppose that $f(x_1, x_2)$ is a superposition of ridge functions with profiles $r_i(t)$, $i = 1, 2, \ldots$ obeying the constraint that, for a certain $s > 1$,*

$$R \equiv \sum_i \|r_i\|_{\dot{B}^s_{1,1}(\mathbf{R})} < \infty; \tag{1.15}$$

that is, for directions θ_i, $i = 1, 2, \ldots$, we have

$$f = \sum_{i=1}^{\infty} r_{i,\theta_i}, \tag{1.16}$$

with the indicated sum necessarily norm-convergent for the $L^\infty(\mathbf{R}^2)$ norm, owing to assumption (1.15). *Then the linear approximation error obeys*

$$\|f - A^{i_1} f\|_{L^\infty(D)} \leq C \cdot 2^{-i_1(s-1)} \cdot R, \qquad i_1 = 1, 2, 3, \ldots.$$

Let $\mathscr{RS}(s, 1, 1, R)$ denote the collection of all functions f arising as superpositions of ridge functions obeying (1.15)–(1.16). The class $\mathscr{RS}(s, 1, 1, R)$ contains, as a subclass, those pure ridge functions $r(x_1 \cos(\theta) + x_2 \sin(\theta))$ with ridge profiles obeying $\|r\|_{\dot{B}^s_{1,1}(\mathbf{R})} \leq R$. On this subclass, Theorem 1.2 shows that the linear ridgelet approximation error is bounded

by $C_2 \cdot 2^{-i_1(s-1)} \cdot R$. In a certain sense, pure ridge functions in a single direction are "as bad as it gets"; superpositions of many directions are no harder.

Consider now an implication of the nonlinear approximation result, Theorem 1.3. Its proof, while initially stated for approximation of pure ridge functions, gives, upon applying the so-called p-triangle inequality as described near (5.10) below, immediate insights for the case of more general functions f.

COROLLARY 1.5. *Suppose that $f(x_1, x_2)$ is a superposition of ridge functions with profiles $r_i(t)$, $i = 1, 2, \ldots$ obeying the constraint*

$$R \equiv \left(\sum_i \|r_i\|^p_{\dot{B}^s_{p,p}(\mathbf{R})} \right)^{1/p} < \infty, \tag{1.17}$$

with $s = 1/p$ and $p \in (0, 1)$. By this we mean that for directions θ_i, $i = 1, 2, \ldots$, we can write

$$f = \sum_{i=1}^{\infty} r_{i,\theta_i}; \tag{1.18}$$

the condition (1.17) *on the ridge profiles guarantees the convergence of this sum in supremum norm over any compact set. Consider the nonlinear approximant f_N defined by setting $f_N \equiv f_\delta$ with f_δ as in* (1.10) *and $N = \bar{N}(\delta)$ (see* (1.9)*). The error obeys*

$$\|f - f_N\|_{L^\infty(D)} \leqslant C \cdot R \cdot N^{-(s-1)}, \qquad N \to \infty,$$

with a constant C that depends only on the disk D and on p.

Let $\mathscr{RS}(s, p, p, R)$ denote the class of all functions f obtainable by superpositions obeying the constraints (1.17)–(1.18). Let $s = 1/p$ and $0 < p < 1$. Now the optimal rate of N-term approximation over the subclass $\mathscr{RS}(s, p, p, R)$ cannot be better than $N^{-(s-1)}$, since—as we saw following (1.12)—that is the best rate of N-term approximation to the subclass of pure ridge functions $r(x_1 \cos(\theta) + x_2 \sin(\theta))$ with profiles obeying $\|r\|_{\dot{B}^s_{p,p}(\mathbf{R})} < R$. But this corollary therefore shows that (f_N) *achieves the same rate of N-term approximation over the entire class $\mathscr{RS}(s, p, p, R)$ as over the subclass consisting of pure ridge functions with profiles $\|r\|_{\dot{B}^s_{p,p}(\mathbf{R})} < R$.* Hence $N^{-(s-1)}$ is the optimal approximation rate over the class $\mathscr{RS}(s, p, p, R)$ and this rate is achieved by (f_N).

In words, the result says that for a certain class of functions f built up nicely out of arbitrarily steered ridge functions, essentially best N-term nonlinear approximations come from thresholded ridgelet expansions.

1.10. *Interpretations*

So far, we have learned two things:

- There is an intimate relation between ridge functions and orthonormal ridgelets. This is illustrated structurally, by the sparse representation formulas (1.3)–(1.4), and the analysis formula (1.7); and also asymptotically, by the approximation rates exhibited in Theorems 1.2 and 1.3.
- Simple algorithms with orthonormal ridgelets might be used in place of complex algorithms based on ridge functions, with equivalent success.

Roughly speaking, while ridge functions are simple to describe, they are hard to use in an effective manner; and while ortho-ridgelets are initially more complicated to understand, they can easily be applied for approximation theory purposes in an effective manner. And the two systems have equivalent approximation properties.

We now elaborate on the difficulties faced by ridge function approximation.

One stream of theoretical results on ridge function approximation is purely existential, and supposes that, one is given directions for an N-term ridge function approximation which are in fact the N optimal directions. This a purely qualitative, abstract approach. To translate this into a practical situation, one hits a roadblock: the assumption of perfect simultaneous global optimal choice of the several direction parameters. For large N this assumption is computationally unrealistic. One doesn't know in general how to numerically find joint minimizers (i.e. jointly optimal directions $\theta_1, ..., \theta_N$).

Another stream of theoretical results is more computationally realistic, and says that we should proceed with sequential choice of directions and apply greedy search methods in order to find the successive directions. While this still requires extensive numerical search, it involves optimization of a single direction at a time, and so is far more computationally realistic. However, the theory then gives only very weak performance guarantees. For example, a function could truly be a superposition of a finite number of nice ridge functions, and yet the rate of convergence of a greedy N-term approximation might be only $1/\sqrt{N}$ [1, 12].

In contrast, the ortho-ridgelet approach proposes a sequence of direct and concrete steps. First, one calculates a fixed discrete set of integrals; then one applies thresholding; finally the terms that survive are used to

provide a partial reconstruction. There is no unrealistic demand for global minimization and no difficult sampling questions. Moreover, as we show here, the convergence rate one gets from simple thresholding is the same as the rate one would get if the underlying optimal directions were known and we were simply trying to build approximations to the underlying ridge profiles. The content of Corollary 1.5 is that if the underlying ridge profiles can existentially be approximated at a fast rate by finite wavelet sums, then the function itself will be well approximated by simple ridgelet thresholding.

In summary, ridgelet-based thresholding is more applicable and concrete than abstract ridge approximation, yet has performance comparable with the optimal performance by abstract procedures. Moreover, in comparison to greedy ridge approximation, ridgelet thresholding gives a far better convergence rate than greedy ridge approximation in certain cases where the optimal rate of ridge approximation is very fast; in such cases, the optimal rate of ridgelet approximation by thresholding can be the same as the rate abstract approximation, while the rate of greedy approximation is only $1/\sqrt{N}$.

1.11. *Application Contexts*

As mentioned earlier, ridge function approximation has been proposed in two contexts.

In the first, associated with tomography [13] and with projection pursuit regression [11], the object is modelled as a superposition of N ridge functions, with complex nonparametric ridge profiles $r_i(t)$—conceivably quite general functions of t.

In the second, associated with Neural Networks [1, 12], the object is modelled as a superposition of N ridge functions, each with a fixed simple ridge shape $\sigma(\cdot)$ (such as a sigmoid), the different profiles differing only in location and scale.

The results quoted above are relevant in both contexts.

On the one hand, we have shown e.g. in Corollary 1.5, that superpositions of N ridge functions having complex nonparametric ridge profiles can be well-approximated by ridgelet thresholding. So ridgelet thresholding is an acceptable approach within the projection pursuit regression setup for ridge approximation.

On the other hand, ridgelet thresholding is very similar to neural nets. We are approximating by (near-) ridge functions (essentially) differing in location, orientation, and scale, so the scheme is like neural network approximation. In that setting, the Corollary 1.5 says that if f is an at-most countable sum $f(x) = \sum_{n=1}^{\infty} \alpha_n \sigma(a_n u_n' x - b_n)$ with $\sum_n |\alpha_n|^p \leqslant R^p$, where the prototype $\sigma(\cdot)$ is smooth and *tends to zero rapidly at* $\pm\infty$, then ridgelet thresholding gives the N-term approximation error $O(N^{-(1/p-1)})$. This is an optimal result under the stated assumptions. For if we even knew the

parameters a_n and u_n and the profile shape $\sigma(\cdot)$ we could not in general have a better rate of N-term approximation than this; and yet ridgelet thresholding works satisfactorily without using any knowledge of p, $\sigma(\cdot)$, or the other parameters. We believe that some extension to the case where σ does not vanish at $\pm\infty$ is also possible.

1.12. *Open Questions*

Ultimately, one would like to know whether or not nonlinear thresholding of the orthonormal ridgelet coefficients is quite generally as effective an approximation method for arbitrary objects f as arbitrary superpositions of sufficiently nice ridge functions. More precisely, supposing f is an object that can be approximated at rate N^{-m} by a sequence of N-term ridge approximants $f_N = \sum_{i=1}^{N} c_{i,N}\rho(a_{i,N}u'_{\theta_{i,N}}x + b_{i,N})$, one would like to know whether or not thresholding of orthonormal ridgelet expansions will give N-term approximations at the rate N^{-m} also.

Corollary 1.5 may be taken as evidence pointing in the direction of an affirmative answer. It shows that for a particular class of objects $\mathscr{RS}(s, p, p, R)$, the worst-case behavior of ridgelet thresholding over the class is not worse than the worst-case behavior of ideal approximation by sums of pure ridge functions. However, one would really like to know whether the two approaches are equivalent on individual functions rather than on functional classes. Although the tools developed in this paper seem very useful to make further progress on such questions, further progress would seem to require new insights as well.

As background, we mention that Candès [4] has shown that dual ridgelets for $L^2(D)$ are almost as good for $L^2(D)$ approximation of individual functions as classical ridge function dictionaries $\{\rho(au'_\theta x + b): (a, u_\theta, b)\}$ with ridge functions based on relatively arbitrary smooth activation functions ρ. Under the hypothesis that approximations to f are built from superpositions of ρ-terms with *bounded coefficients* he has been able to show that the rate of approximation by N-term *dual ridgelet* expansions is at most logarithmically worse than the rate of approximation by ridge functions with smooth activation units. Unfortunately, little of an explicit nature is known about the elements of the approximating dual ridgelet frames, beyond the fact that they are certainly not ridge functions. But the result is very suggestive.

Finally, much of the interest in ridge function approximation has to do with high dimensions [5], so extensions of these results beyond dimension two would be of interest. The underlying tool used here—the ortho-ridgelet basis—has been generalized to high dimensions, producing, however, not an orthobasis, but instead a tight frame in all dimensions greater than 2 [10]. It seems a reasonable goal to seek generalizations of the results here for dimensions >2.

2. PRELIMINARIES

2.1. *Orthonormal Ridgelets and Radon Transform*

We begin by briefly quoting some material from [9]. 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 $\mathscr{L}_{(\theta, t)}$, expressed using the Dirac mass δ as

$$(Rf)(t, \theta) = \int f(x)\, \delta(x_1 \cos\theta + x_2 \sin\theta - t)\, dx, \tag{2.1}$$

where we permit $\theta \in [0, 2\pi)$ and $t \in \mathbf{R}$. Note that Rf has the antipodal symmetry

$$(Rf)(-t, \theta + \pi) = (Rf)(t, \theta). \tag{2.2}$$

We adopt the convention that F (and G and variants) typically will denote a function on $\mathbf{R} \times [0, 2\pi)$ obeying the same antipodal symmetry:

$$F(-t, \theta + \pi) = F(t, \theta). \tag{2.3}$$

To create a space of such objects, we let $[\ ,\ ]$ denote the pairing

$$[F, G] = \frac{1}{4\pi} \int_0^{2\pi} \int_{-\infty}^{\infty} F(t, \theta)\, \bar{G}(t, \theta)\, dt\, d\theta, \tag{2.4}$$

and by $L^2(dt\, d\theta)$ norm we mean $\|F\|^2 = [F, F]$. Let $\mathscr{R}$ be the closed subspace of $L^2(dt\, d\theta)$ of functions F obeying (2.3). Let $P_{\mathscr{R}} F$ be the orthoprojector from $L^2(dt\, d\theta)$ onto $\mathscr{R}$, defined by

$$(P_{\mathscr{R}} F)(t, \theta) = (F(t, \theta) + F(-t, \theta + \pi))/2. \tag{2.5}$$

Define the operator of reflection of functions of one variable $(Tf)(t) = f(-t)$ and the operator of translation by half a period by $(Sg)(\theta) = g(\theta + \pi)$. Note that the space $\mathscr{R}$ consists of objects invariant under $T \otimes S$; (2.3) can be rewritten $(T \otimes S) F = F$. In fact, $P_{\mathscr{R}} = (I + T \otimes S)/2$. Set now, for $j, k \in \mathbf{Z}$, and $i \geqslant i_0$, $\ell = 0, ..., 2^{i-1} - 1$, $\varepsilon \in \{0, 1\}$

$$W_\lambda(t, \theta) = P_{\mathscr{R}}(\psi_{j,k} \otimes w^\varepsilon_{i,\ell}), \tag{2.6}$$

where $\lambda = (j, k; i, \ell, \varepsilon)$. For later reference, we spell this out:

$$W_\lambda(t, \theta) = (\psi_{j,k}(t)\, w^\varepsilon_{i,\ell}(\theta) + \psi_{j,k}(-t)\, w^\varepsilon_{i,\ell}(\theta + \pi))/2. \tag{2.7}$$

It was shown in [9] that the W_λ provide an orthobasis for $\mathscr{R}$. In order to obtain orthonormality with respect to the scalar product $[\ ,\ ]$, a particular

normalization was imposed. In that normalization, $\|\psi_{j,k}\|_{L^2(\mathbf{R})} = \sqrt{2}$, and $\|w^{\varepsilon}_{i,\ell}\| = 2\sqrt{\pi}$. In a sense, the $(W_\lambda : \lambda \in \Lambda)$ constitute a "tensor wavelet basis which has been antipodally symmetrized".

We define the adjoint of the Radon transform so that for all sufficiently nice $G \in \mathscr{R}$ and all sufficiently nice $f \in L^2(dx)$,

$$[Rf, G] = \langle f, R^+ G \rangle, \tag{2.8}$$

which leads to

$$(R^+ G)(x) = \frac{1}{4\pi} \int_0^{2\pi} G(x_1 \cos\theta + x_2 \sin\theta, \theta)\, d\theta. \tag{2.9}$$

Define the Riesz order-1/2 fractional differentiation operator Δ^+ and also the order-1/2 fractional integration operator Δ^- by the unified formula

$$(\Delta^{\pm} f)(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{it\omega} \hat{f}(\omega)\, |\omega|^{\pm 1/2}\, d\omega. \tag{2.10}$$

These unbounded operators are well-defined on functions which are sufficiently smooth [formally, the domain $\mathscr{D}(\Delta^+) = \{f: \int_{-\infty}^{\infty} |\hat{\rho}(\omega)|^2\, |\omega|\, d\omega < \infty\}$] or sufficiently oscillatory [formally, the domain $\mathscr{D}(\Delta^-) = \{f: \int_{-\infty}^{\infty} |\hat{f}(\omega)|^2\, |\omega|^{-1}\, d\omega < \infty\}$]; in particular, they are well-defined on every 1-D Meyer wavelet $\psi_{j,k}$, owing to $\mathrm{supp}(\hat{\psi}_{j,k}) \subset \{\omega: |\omega| \in [\frac{2}{3}\pi 2^j, \frac{8}{3}\pi 2^j]\}$. Moreover, on the appropriate domains, they are self-adjoint; and on the appropriate domains they act as inverses of each other.

Set now, for $\lambda \in \Lambda$,

$$\tau_\lambda = (\Delta^+ \otimes I)\, W_\lambda, \qquad \lambda \in \Lambda. \tag{2.11}$$

For example,

$$\tau_{(j,k;i,\ell,1)} = (\Delta^+ \psi_{j,k} \otimes w^1_{i,\ell} + \Delta^+ T\psi_{j,k} \otimes S w^1_{i,\ell})/2.$$

A useful remark is that $\Delta^{\pm} T = T\Delta^{\pm}$ on the appropriate domains. Of course, in terms of the functions $\psi^+_{j,k}$ of the introduction, we have $\psi^+_{j,k} = \Delta^+ \psi_{j,k}$.

The key formula we need for this paper is

$$\rho_\lambda = R^+[(\Delta^+ \otimes I)\, W_\lambda] = R^+[\tau_\lambda]. \tag{2.12}$$

The operator R^+ is typically called back-projection in the tomography literature. This formula says that an orthonormal ridgelet is the *back-projection of a fractionally-differentiated wavelet which has been anti-podally-symmetrized.*

2.2. *Special Ridge Functions*

We now consider the relationship (1.3)–(1.4) between special ridge functions r^γ and ridgelets ρ^γ. Formally, this is a simple matter, based immediately on (2.12) and the definition of the W_λ. With δ_{θ_0} denoting the Dirac mass at $\theta_0 \in [0, 2\pi)$,

$$\begin{aligned} r^\gamma(x) &= \psi^+_{j,k}(x_1 \cos\theta_0 + x_2 \sin\theta_0) \\ &= 4\pi \cdot R^+[(\psi^+_{j,k} \otimes \delta_{\theta_0} + \psi^+_{j,1-k} \otimes \delta_{\theta_0+\pi})/2] \\ &= 4\pi \cdot R^+[(\Delta^+ \otimes I)(\psi_{j,k} \otimes \delta_{\theta_0} + \psi_{j,1-k} \otimes \delta_{\theta_0+\pi})/2]. \end{aligned} \tag{2.13}$$

In a moment we will argue that

$$4\pi \cdot (\psi_{j,k} \otimes \delta_{\theta_0} + \psi_{j,1-k} \otimes \delta_{\theta_0+\pi})/2 = \sum a^\gamma_\lambda W_\lambda. \tag{2.14}$$

Combining (2.13) and (2.14), we have

$$\begin{aligned} r^\gamma &= R^+\left[(\Delta^+ \otimes I) \sum a^\gamma_\lambda W_\lambda \right] \\ &= R^+\left[\sum a^\gamma_\lambda (\Delta^+ \otimes I)\, W_\lambda \right] \\ &= \sum a^\gamma_\lambda R^+[(\Delta^+ \otimes I)\, W_\lambda] \\ &= \sum a^\gamma_\lambda \rho_\lambda, \end{aligned}$$

completing the formal argument for (1.4).

The argument for (2.14) is as follows. Let $\tilde{w}^e_{i,\ell}$ denote the basis element $w^e_{i,l}$ rescaled to unit $L^2[0, 2\pi]$ norm. In the sense of distributions, we have the identity

$$\delta_{\theta_0} = \sum_{\ell=0}^{2^{i_0}-1} \langle \delta_{\theta_0}, \tilde{w}^0_{i_0,\ell} \rangle\, \tilde{w}^0_{i_0,\ell} + \sum_{i=i_0}^{\infty} \sum_{\ell=0}^{2^i-1} \langle \delta_{\theta_0}, \tilde{w}^1_{i,\ell} \rangle\, \tilde{w}^1_{i,\ell};$$

note here the expanded range of summation in ℓ, the range being $0 \leqslant \ell < 2^i$ rather than $0 \leqslant \ell < 2^{i-1}$ as in the definition of the set Λ from Section 1. In terms of the basis $w^e_{i\ell}$ *without* rescaling,

$$4\pi \cdot \delta_{\theta_0} = \sum_{\ell=0}^{2^{i_0}-1} \langle \delta_{\theta_0}, w^0_{i_0,\ell} \rangle\, w^0_{i_0,\ell} + \sum_{i=i_0}^{\infty} \sum_{\ell=0}^{2^i-1} \langle \delta_{\theta_0}, w^1_{i,\ell} \rangle\, w^1_{i,\ell}. \tag{2.15}$$

Hence, again in the sense of distributions,

$$
\begin{aligned}
&4\pi\cdot(\psi_{j,k}\otimes\delta_{\theta_0}+\psi_{j,1-k}\otimes\delta_{\theta_0+\pi})/2\\
&\quad=\tfrac{1}{2}\sum_{\ell=0}^{2^{i_0}-1}\langle\delta_{\theta_0},w^0_{i_0,\ell}\rangle\,\psi_{j,k}\otimes w^0_{i_0,\ell}+\tfrac{1}{2}\sum_{i=i_0}^{\infty}\sum_{\ell=0}^{2^i-1}\langle\delta_{\theta_0},w^1_{i,\ell}\rangle\,\psi_{j,k}\otimes w^1_{i,\ell}\\
&\qquad+\tfrac{1}{2}\sum_{\ell=0}^{2^{i_0}-1}\langle\delta_{\theta_0+\pi},w^0_{i_0,\ell}\rangle\,\psi_{j,1-k}\otimes w^0_{i_0,\ell}\\
&\qquad+\tfrac{1}{2}\sum_{i=i_0}^{\infty}\sum_{\ell=0}^{2^i-1}\langle\delta_{\theta_0+\pi},w^1_{i,\ell}\rangle\,\psi_{j,1-k}\otimes w^1_{i,\ell}.
\end{aligned}
$$

Rearranging the first term of the sum gives a sum over half as many terms:

$$
\psi_{j,k}\otimes\sum_{\ell=0}^{2^{i_0-1}-1}(\langle\delta_{\theta_0},w^0_{i_0,\ell}\rangle\,w^0_{i_0,\ell}+\langle\delta_{\theta_0},w^0_{i_0,\ell+2^{i_0-1}}\rangle\,w^0_{i_0,\ell+2^{i_0-1}})/2.
$$

Rearranging the third term of that same sum gives, similarly,

$$
\psi_{j,1-k}\otimes\sum_{\ell=0}^{2^{i_0-1}-1}(\langle\delta_{\theta_0+\pi},w^0_{i_0,\ell}\rangle\,w^0_{i_0,\ell}+\langle\delta_{\theta_0+\pi},w^0_{i_0,\ell+2^{i_0-1}}\rangle\,w^0_{i_0,\ell+2^{i_0-1}})/2.
$$

We now notice that $\langle\delta_{\theta_0+\pi},w^0_{\ell+2^{i_0-1}}\rangle=\langle\delta_{\theta_0},w^0_{i_0,\ell}\rangle$, so summing the last two displays and rearranging subexpressions, we obtain

$$
\begin{aligned}
&\sum_{\ell=0}^{2^{i_0-1}-1}\langle\delta_{\theta_0},w^0_{i_0,\ell}\rangle(\psi_{j,k}\otimes w^0_{i_0,\ell}+\psi_{j,1-k}\otimes w^0_{i_0,\ell+2^{i_0-1}})/2\\
&\qquad+\sum_{\ell=0}^{2^{i_0-1}-1}\langle\delta_{\theta_0+\pi},w^0_{i_0,\ell}\rangle(\psi_{j,1-k}\otimes w^0_{i_0,\ell}+\psi_{j,k}\otimes w^0_{i_0,\ell+2^{i_0-1}})/2.
\end{aligned}
$$

Now from $\psi_{j,k}(-t)=\psi_{j,1-k}(t)$ and $\langle\delta_{\theta_0+\pi},w^0_{i_0,\ell}\rangle=w^0_{i_0,\ell}(\theta_0+\pi)$, we can rearrange this to

$$
\sum_{\ell=0}^{2^{i_0-1}-1}w^0_{i_0,\ell}(\theta_0)\cdot W_{(j,k;i_0,\ell,0)}+\sum_{\ell=0}^{2^{i_0-1}-1}w^0_{i_0,\ell}(\theta_0+\pi)\cdot W_{(j,1-k;i_0,\ell,0)}.
$$

Obviously, similar relationships hold for sums involving terms $w^1_{i,\ell}$ at indices $i>i_0$, and so

$$
\begin{aligned}
&4\pi\cdot(\psi_{j,k}\otimes\delta_{\theta_0}+\psi_{j,1-k}\otimes\delta_{\theta_0+\pi})/2\\
&\quad=\sum_{\ell=0}^{2^{i_0-1}-1}w^0_{i_0,\ell}(\theta_0)\;W_{(j,k;i_0,\ell,0)}+w^0_{i_0,\ell}(\theta_0+\pi)\;W_{(j,1-k;i_0,\ell,0)}\\
&\qquad+\sum_{i=i_0}^{\infty}\sum_{\ell=0}^{2^{i-1}-1}w^1_{i,\ell}(\theta_0)\;W_{(j,k;i,\ell,1)}+w^1_{i,\ell}(\theta_0+\pi)\;W_{(j,1-k;i,\ell,1)}.
\end{aligned}
$$

Comparing with the definition of W_λ, this gives (2.14), and completes the formal motivation of (1.7).

To attach a rigorous meaning to (1.4), we work with a weighted L^∞-norm $\|f\|_{\infty,-2\sigma} = \sup_{x\in\mathbf{R}^2} |f(x)|\,(1+|x|^2)^{-\sigma}$, $\sigma \in \{3, 4, 5, ...\}$; this down-weights x's far from 0. We write $L^\infty_{-2\sigma}$ for the vector space of f with finite $\|f\|_{\infty,-2\sigma}$. Let Λ^{i_1} denote the collection $\{\lambda: j, i < i_1\}$ of indices λ that are index-limited to i, j coarser than i_1. Consider the index-limited partial sum approximation $r^{(i_1)} = \sum_{\Lambda^{i_1}} a^\gamma_\lambda \rho_\lambda$.

LEMMA 2.1. *Let $\sigma \in \{3, 4, 5, ...\}$. The identity*

$$r^\gamma = \sum_\lambda a^\gamma_\lambda \rho_\lambda$$

makes sense in $L^\infty_{-2\sigma}$: the approximation $r^{(i_1-1)}$ is norm-convergent to the left-hand side in $L^\infty_{-2\sigma}$ as $i_1 \to \infty$.

Proof. For $i_2 > i_1$, put

$$E_{i_1,i_2} = \|r^{(i_1)} - r^{(i_2)}\|_{L^\infty_{-2\sigma}}.$$

Then, for $\partial\Lambda_{i_1,i_2} = \Lambda^c_{i_2} \setminus \Lambda^c_{i_1}$, $(i_2 > i_1)$,

$$\begin{aligned} E_{i_1,i_2} &= \left\| \left(\sum_{\lambda\in\partial\Lambda_{i_1,i_2}} a^\gamma_\lambda \rho_\lambda(\cdot) \right) (1+|\cdot|^2)^{-\sigma} \right\|_{L^\infty} \\ &\leqslant \sum_{\lambda\in\partial\Lambda_{i_1,i_2}} |a^\gamma_\lambda|\, \|\rho_\lambda(\cdot)\cdot(1+|\cdot|^2)^{-\sigma}\|_{L^\infty} \\ &\leqslant \sum_{\lambda\in\Lambda^c_{i_1}} |a^\gamma_\lambda|\, \|\rho_\lambda(\cdot)\cdot(1+|\cdot|^2)^{-\sigma}\|_{L^\infty}. \end{aligned}$$

Now pick $m > 1$. By rapid decay of $w^\varepsilon_{i,\ell}(\theta)$, there is C_m so that if $\lambda = (j_0, k_0; i, \ell, \varepsilon)$ or $(j_0, 1-k_0; i, \ell, \varepsilon)$, for some i, ℓ, then

$$|a^\gamma_\lambda| \leqslant C_m 2^{i/2}((1+2^i\,|\theta_0 - \theta_{i,\ell}|)^{-m} + (1+2^i\,|\theta_0 + \pi - \theta_{i,\ell}|)^{-m}).$$

Now as $i_2 > i_1 \geqslant i_0$, $\varepsilon(\lambda) = 1$ at all terms occuring in E_{i_1,i_2}. We can apply (2.16) of the Lemma following, with $0 < d \leqslant 2(\sigma-2)$, and we can use the observation that $|a^\gamma_\lambda| = 0$ unless $j = j_0$, getting

$$\begin{aligned} E_{i_1,i_2} &\leqslant C(j_0) \sum_{i\geqslant i_1} 2^{-i(d+1/2)} 2^{i/2} \\ &\quad \times \sum_\ell [(1+2^i\,|\theta_0 - \theta_{i,\ell}|)^{-m} + (1+2^i\,|\theta_0 + \pi - \theta_{i,\ell}|)^{-m}] \\ &= C'(j_0) \sum_{i\geqslant i_1} 2^{-id} \leqslant C''(j_0)\, 2^{-i_1 d}. \end{aligned}$$

As $E_{i_1, i_2} \leqslant C''(j_0)\, 2^{-i_1 d}$ we conclude that $(r^{(i_1)})_{i_1}$ is a Cauchy sequence in $L^{\infty}_{-2\sigma}$ norm.

We now show that the limit of this sequence is r^{γ}. The key observation is

$$\sum_{\ell=0}^{2^{i_0}-1} \langle \tilde{w}^0_{i_0,\ell}, \delta_{\theta_0} \rangle\, \tilde{w}^0_{i_0,\ell} + \sum_{i=i_0}^{i_1-1} \sum_{\ell=i_0}^{2^i-1} \langle \tilde{w}^1_{i,\ell}, \delta_{\theta_0} \rangle\, \tilde{w}^1_{i,\ell} = \sum_{\ell=0}^{2^{i_1}-1} \langle \tilde{w}^0_{i_1,\ell}, \delta_{\theta_0} \rangle\, \tilde{w}^0_{i_1,\ell}$$

which is merely the usual "rewriting rule" in the theory of multiresolution analysis, the rule for converting from a monoscale representation by scaling coefficients at fine level i_1, to a multiscale representation by scaling coefficients at coarse level $i_0 < i_1$ and wavelet coefficients at levels $i_0 \leqslant i < i_1$. Note that this relationship is habitually stated in terms of the orthonormal functions $\tilde{w}^e_{i,\ell}$, but would remain equally valid in terms of $w^e_{i,\ell}$. Let $\tilde{P}_{i_1}\delta_{\theta_0} = \sum_{\ell=0}^{2^{i_1}-1} \langle \tilde{w}^0_{i_1,\ell}, \delta_{\theta_0} \rangle\, \tilde{w}^0_{i_1,\ell}$, and $P_{i_1}\delta_{\theta_0} = \sum_{\ell=0}^{2^{i_1}-1} \langle w^0_{i_1,\ell}, \delta_{\theta_0} \rangle\, w^0_{i_1,\ell}$. Then $P = 4\pi \cdot \tilde{P}$. It follows immediately from the rewriting rule for P and the definition of a^{γ}_{λ} that

$$\begin{aligned} \sum_{\lambda \in \Lambda^{i_1}} a^{\gamma}_{\lambda} \tau_{\lambda} &= (\psi^+_{j_0,k_0} \otimes P_{i_1}\delta_{\theta_0} + \psi^+_{j_0,1-k_0} \otimes P_{i_1}\delta_{\theta_0+\pi})/2 \\ &= 4\pi \cdot (\psi^+_{j_0,k_0} \otimes \tilde{P}_{i_1}\delta_{\theta_0} + \psi^+_{j_0,1-k_0} \otimes \tilde{P}_{i_1}\delta_{\theta_0+\pi})/2. \end{aligned}$$

Now for the finite sum $r^{(i_1)}$, we have

$$\begin{aligned} r^{(i_1)} &= \sum_{\lambda \in \Lambda^{i_1}} a^{\gamma}_{\lambda} \rho_{\lambda} = R^+ \left[\sum_{\lambda \in \Lambda^{i_1}} a^{\gamma}_{\lambda} \tau_{\lambda} \right] \\ &= R^+[(\psi^+_{j_0,k_0} \otimes P_{i_1}\delta_{\theta_0} + \psi^+_{j_0,1-k_0} \otimes P_{i_1}\delta_{\theta_0+\pi})/2] \\ &= 4\pi \cdot R^+[(\psi^+_{j_0,k_0} \otimes \tilde{P}_{i_1}\delta_{\theta_0} + \psi^+_{j_0,1-k_0} \otimes \tilde{P}_{i_1}\delta_{\theta_0+\pi})/2]. \end{aligned}$$

Put $r_{j,k,\theta}(x) \equiv \psi^+_{j,k}(x_1 \cos(\theta) + x_2 \sin(\theta))$; fix x and view θ as the variable. From the definition (2.9) of R^+,

$$r^{(i_1)}(x) = 4\pi \cdot \frac{1}{4\pi} \int_0^{2\pi} (r_{j_0,k_0,\theta}(x)\, \tilde{P}_{i_1}\delta_{\theta_0}(\theta) + r_{j_0,1-k_0,\theta}(x)\, \tilde{P}_{i_1}\delta_{\theta_0+\pi})/2\, d\theta.$$

We note that whenever $g(\theta)$ is a continuous function on $[0, 2\pi)$,

$$\int_0^{2\pi} (\tilde{P}_{i_1}\delta_{\theta_0})(\theta)\, g(\theta)\, d\theta \to g(\theta_0) \qquad \text{as} \quad i_1 \to \infty.$$

For fixed x, $r_{j,k,\theta}(x)$ is uniformly continuous in θ. Hence as $i_1 \to \infty$,

$$r^{(i_1)}(x) \to (r_{j_0,k_0,\theta_0}(x) + r_{j_0,1-k_0,\theta_0+\pi}(x))/2 = r_{j_0,k_0,\theta_0}(x) = r^{\gamma}(x).$$

This pointwise convergence proves that $r^{(i_1)} \to r^{\gamma}$ in $L^{\infty}_{-2\sigma}$ norm. ∎

LEMMA 2.2. *For $\sigma \geqslant 3$ and $d = 1, ..., 2(\sigma-2)$, for $\varepsilon = 1$,*

$$\|\rho_{\lambda}(\cdot)(1+|\cdot|^2)^{-\sigma}\|_{L^{\infty}} \leqslant C(j)\, 2^{-i(d+1/2)}, \qquad i \geqslant \max(i_0, j(\lambda)). \quad (2.16)$$

Proof. Let $K(x) = (1+|x|^2)^{-\sigma}$; then $K \in L^1$ provided $\sigma > 1$, and $x_i^{\ell} K \in L^1$ provided $\sigma > 1 + \ell/2$. It follows that $\hat{K}(\xi) \in C^{2(\sigma-2)}$ and that $\hat{K} \in C^{\infty}(\mathbf{R}^2 \backslash \{0\})$. In fact $\hat{K}(\xi)$, along with its derivatives, is of rapid decay as $|\xi| \to \infty$.

Employing this terminology, (2.16) can be estimated by

$$\|\rho_{\lambda} \cdot K\|_{L^{\infty}} \leqslant \|\hat{\rho}_{\lambda} * \hat{K}\|_{L^1}. \quad (2.17)$$

Now as $\hat{\rho}_{\lambda}$ is supported in an annulus Γ_j, we have

$$(\hat{\rho}_{\lambda} * \hat{K})(\xi) = \iint_{\Gamma_j} \hat{\rho}_{\lambda}(\xi')\, \hat{K}(\xi - \xi')\, d\xi';$$

making a polar coordinate transformation $\xi \leftrightarrow (r, \theta)$, Γ_j transforms into a rectangle $[R_j, R_{j+1}] \times [0, 2\pi]$, and this becomes

$$\int_{R_j}^{R_j} \int_0^{2\pi} (\hat{\psi}_{j,k}(r)\, w^{e}_{i,\ell}(\theta) + \hat{\psi}_{j,1-k}(r)\, w^{e}_{i,\ell+2^{i-1}}(\theta))/2$$
$$\times \hat{K}(\xi - (r\cos\theta, r\sin\theta))\, d\theta\, r^{1/2}\, dr.$$

For $d \in \{1, 2, ..., 2(\sigma-2)\}$ set

$$\tilde{K}^{(d)}(r, \theta; \xi) = \left(\frac{\partial}{\partial\theta}\right)^d \hat{K}(\xi - (r\cos\theta, r\sin\theta)),$$

and note that, with $\hat{K}^{b,c} \equiv (\frac{\partial}{\partial\xi_1})^b (\frac{\partial}{\partial\xi_2})^c \hat{K}(\xi)$,

$$\tilde{K}^{(d)}(r, \theta; \xi) = \sum_{b=0}^{d} \sum_{c=0}^{d} \hat{K}^{b,c}(\xi - (r\cos\theta, r\sin\theta)) \cdot P_{b,c}(r\cos\theta, r\sin\theta),$$

where each $P_{b,c}(x_1, x_2)$ is a polynomial of degree $\leqslant d$. Now

$$\left\| \left(\frac{\partial}{\partial\xi_1}\right)^b \left(\frac{\partial}{\partial\xi_2}\right)^c \hat{K}(\xi) \right\|_{L^1} \leqslant C_d \qquad 0 \leqslant c \leqslant b \leqslant d < 2(\sigma-2).$$

Hence, with $C(j) = \sup_{0 \leqslant b,\, c \leqslant d} \sup_{x \in \Gamma_j} |P_{b,\, c}(x_1, x_2)|$,

$$\|\tilde{K}^{(d)}(\cdot,\ \cdot\,;\xi)\|_{L^1(dr\, d\theta)} \leqslant C_d \cdot C(j).$$

Now for each fixed $r > 0$, if $\varepsilon = 1$, $w^{\varepsilon}_{i,\,\ell}$ has d-fold primitive, and

$$\int w^{\varepsilon}_{i,\,\ell}(\theta)\, \hat{K}(\xi - (r\cos\theta, r\sin\theta))\, d\theta$$

$$= (-1)^d \int (w^{\varepsilon}_{i,\,\ell})^{(-d)}(\theta)\, \tilde{K}^{(d)}(r, \theta; \xi)\, d\theta.$$

Hence

$$(\hat{\rho}_\lambda * \hat{K})(\xi) = \int_{R_j}^{R_{j+1}} \hat{\psi}_{j,\,k}(r) \left(\int_0^{2\pi} (w^{\varepsilon}_{i,\,\ell})^{(-d)}(\theta)\, \tilde{K}^{(d)}(r, \theta\ \ \xi)\, d\theta \right) r^{1/2}\, dr$$

$$+ \int_{R_j}^{R_{j+1}} \hat{\psi}_{j,\,1-k}(r) \left(\int_0^{2\pi} (w^{\varepsilon}_{i,\,\ell})^{(-d)}(\theta + \pi)\, \tilde{K}^{(d)}(r, \theta; \xi)\, d\theta \right) r^{1/2}\, dr,$$

and, by Minkowski,

$$\|\hat{\rho}_\lambda * \hat{K}\|_{L^1} \leqslant \sup_{\xi \in \Gamma_j} \|\tilde{K}^{(d)}(\cdot,\ \cdot\,;\xi)\|_{L^2(dr\, d\theta)}$$

$$\times \left(\int_{R_j}^{R_{j+1}} \int_0^{2\pi} |\hat{\psi}_{j,\,k}(r)|\, r^{1/2}\, |(w^{\varepsilon}_{i,\,\ell})^{(-d)}(\theta)|\, d\theta\, dr \right.$$

$$\left. + \int_{R_j}^{R_{j+1}} \int_0^{2\pi} |\hat{\psi}_{j,\,1-k}(r)|\, r^{1/2}\, |(w^{\varepsilon}_{i,\,\ell})^{(-d)}(\theta + \pi)|\, d\theta\, dr \right) \Big/ 2$$

$$= C_d \cdot \int_0^{2\pi} |(w^{\varepsilon}_{i,\,\ell})^{(-d)}(\theta)|\, d\theta$$

$$= C_d \cdot 2^{-i(d+1/2)}. \quad \blacksquare$$

3. ANALYSIS OF RIDGE FUNCTIONS

We now turn to Theorem 1.1 of the Introduction, giving a very simple formula for ridgelet coefficients of a ridge function—at least for ridge functions r_θ with profile $r \in B^1_{1,\,1}(\mathbf{R})$.

3.1. *Preliminaries*

LEMMA 3.1. *Let $r(t)$ belong to the Bump algebra—Besov space $\dot{B}^1_{1,1}(\mathbf{R})$ —and vanish at $\pm\infty$. Then r is bounded, and there exists an enumeration (j_n, k_n) of the indices (j,k) so that the sequence of finite sums*

$$r^{(N)} = \sum_{n=1}^{N} c_{j_n,k_n}\psi^+_{j_n,k_n}$$

converges in $L^\infty(\mathbf{R})$ norm to r.

This will be proven in the Appendix. The proof rests on the following lemma, also established in the Appendix.

LEMMA 3.2. *Let $r(t) \in \dot{B}^s_{1,1}(\mathbf{R})$, $s \geqslant 1$, and vanish at $\pm\infty$. There exist coefficients $(c_{j,k}: j,k \in \mathbf{Z})$ so that*

$$r = \sum c_{j,k}\psi^+_{j,k},$$

where the right side converges in norm to $r(t)$ in $L^\infty(\mathbf{R})$; the coefficients obey

$$\sum |c_{j,k}|\, 2^{js} < C \cdot \|r\|_{\dot{B}^s_{1,1}(\mathbf{R})}. \tag{3.1}$$

3.2. *Proof of Theorem* 1.1.

Let $\gamma = (j_0, k_0, \theta_0)$ and a^γ_λ as in (1.3). Let $r^{(i_1)} = \sum_{\Lambda^{i_1}} a^\gamma_\lambda \rho_\lambda$. Now for fixed $m > 1$, $r^{(i_1)}$ is $L^\infty_{-2\sigma}$-norm convergent to r as $i_1 \to \infty$. On the other hand, each $\rho_\lambda \in L^1_{2\sigma}$, where $L^1_{2\sigma}$ is the space of functions on $\mathbf{R}^2$ normed by $\|f\|_{L^1_{2\sigma}} = \int |f(x)|\,(1+|x|^2)^\sigma\, dx$. As $L^1_{2\sigma} \subset (L^\infty_{-2\sigma})^*$, it follows that each $\langle \rho_\lambda, \cdot \rangle$ defines a bounded linear functional on $L^\infty_{-2\sigma}$. Hence

$$\lim_{i_1 \to \infty} \langle r^{(i_1)}, \rho_\lambda \rangle = \langle r^\gamma, \rho_\lambda \rangle \qquad \forall \lambda \in \Lambda.$$

Now each $r^{(i_1)}$ is a finite sum of ρ_λ, so from orthogonality $\langle \rho_\lambda, \rho_{\lambda'} \rangle = \delta_{\lambda\lambda'}$ we get

$$\langle r^{(i_1)}, \rho_\lambda \rangle = \begin{cases} 0 & \lambda \notin \Lambda^{i_1} \\ a^\gamma_\lambda & \lambda \in \Lambda^{i_1}. \end{cases}$$

It follows that

$$\langle r^\gamma, \rho_\lambda \rangle = a^\gamma_\lambda \qquad \forall \lambda \in \Lambda. \tag{3.2}$$

By the orthogonality $\langle \psi^+_{j,k}, \psi^-_{j',k'} \rangle = 2 \cdot \delta_{jj'} \delta_{kk'}$, this can be written as

$$\langle r^\gamma, \rho_\lambda \rangle = (\langle \psi^+_{j_0,k_0}, \psi^-_{j,k} \rangle \cdot w^{\varepsilon}_{i,\ell}(\theta_0) + \langle \psi^+_{j_0,k_0}, \psi^-_{j,1-k} \rangle \cdot w^{\varepsilon}_{i,\ell}(\theta_0 + \pi))/2 \quad (3.3)$$

which proves (1.7) in the special case where r_θ is a special ridge function, i.e. for ridge profile $r = \psi^+_{j_0,k_0}$. We easily get the full result (1.7) from this special case. Let r be as in the statement of the Theorem, and let $(j_n, k_n)_n$ be the corresponding enumeration of Lemma 3.1. Define the composite ridge profile $r_N = \sum_{n=1}^N c_n \psi_{j_n,k_n}$ and corresponding ridge function $r_{N,\theta_0} = \sum_{n=1}^N c_n r^{\gamma_n}$, with $\gamma_n = (j_n, k_n, \theta_0)$, and note that (1.7) follows immediately for all such finite composites, simply from superposing (3.3):

$$\begin{aligned}
\langle r_{N,\theta_0}, \rho_\lambda \rangle &= \left\langle \sum_{n=1}^N c_n r^{\gamma_n}, \rho_\lambda \right\rangle \\
&= \sum_{n=1}^N c_n \cdot \langle r^{\gamma_n}, \rho_\lambda \rangle \\
&= \sum_{n=1}^N c_n \cdot (\langle \psi^+_{j_n,k_n}, \psi^-_{j,k} \rangle \cdot w^{\varepsilon}_{i,\ell}(\theta_0) \\
&\quad + \langle \psi^+_{j_n,k_n}, \psi^-_{j,1-k} \rangle \cdot w^{\varepsilon}_{i,\ell}(\theta_0 + \pi))/2 \\
&= \left(\left\langle \sum_{n=1}^N c_n \psi^+_{j_n,k_n}, \psi^-_{j,k} \right\rangle \cdot w^{\varepsilon}_{i,\ell}(\theta_0) \right. \\
&\quad \left. + \left\langle \sum_{n=1}^N c_n \psi^+_{j_n,k_n}, \psi^-_{j,1-k} \right\rangle \cdot w^{\varepsilon}_{i,\ell}(\theta_0 + \pi) \right) \Big/ 2 \\
&= (\langle r_N, \psi^-_{j,k} \rangle \cdot w^{\varepsilon}_{i,\ell}(\theta_0) + \langle r_N, \psi^-_{j,1-k} \rangle \cdot w^{\varepsilon}_{i,\ell}(\theta_0 + \pi))/2.
\end{aligned}$$

But

$$r_{N,\theta_0} \to r_{\theta_0} \quad \text{in } L^\infty(\mathbf{R}^2) \quad \text{and} \quad r_N \to r \quad \text{in } L^\infty(\mathbf{R}) \qquad \text{as} \quad N \to \infty.$$

As each $\rho_\lambda \in L^1(\mathbf{R}^2) \subset (L^\infty(\mathbf{R}^2))^*$, and as $\psi^-_{j,k} \in L^1(\mathbf{R}) \subset (L^\infty(\mathbf{R}))^*$,

$$\begin{aligned}
\langle r_{N,\theta_0}, \rho_\lambda \rangle &\to \langle r_{\theta_0}, \rho_\lambda \rangle \qquad \forall \lambda \in \Lambda, \\
\langle r_N, \psi^-_{j,k} \rangle &\to \langle r, \psi^-_{j,k} \rangle \qquad \forall j, k \in \mathbf{Z}, \qquad \text{as} \quad N \to \infty.
\end{aligned}$$

So (1.7) follows as claimed.

4. SYNTHESIS OF RIDGE FUNCTIONS

In this section we prove Theorem 1.2.

4.1. *Preliminaries*

We now consider the formal association

$$r_{\theta_0} \sum_{\Lambda} \langle r_{\theta_0}, \rho_\lambda \rangle \rho_\lambda$$

and develop a rigorous meaning for this expression. The coefficients $\langle r_{\theta_0}, \rho_\lambda \rangle$ may in certain cases grow with increasing $i = i(\lambda)$, so it is not immediately apparent that this can be done.

We localize attention to the unit disc $D = \{x \in \mathbf{R}^2 : |x| \leqslant 1\}$, and consider the approximation operator

$$A^{i_1} f = \sum_{\lambda \in \Lambda^{i_1}} \langle f, \rho_\lambda \rangle \rho_\lambda$$

which can be shown to make sense for nice ridge functions $f = r_{\theta_0}$.

LEMMA 4.1. *Let $r \in \dot{B}^1_{1,1}$, and r vanish at $\pm\infty$. Then for each $i_1 \geqslant i_0$ the sum $\sum_{\Lambda^{i_1}} \langle r_\theta, \rho_\lambda \rangle \rho_\lambda$ is absolutely convergent for the $L^\infty(D)$ norm.*

Proof. Decompose Λ^{i_1} into layers $L^i = \Lambda^i \backslash \Lambda^{i-1}$; for $i_1 > i_0$ put

$$\Lambda^{i_1} = \Lambda^{i_0} \cup L^{i_0+1} \cup \cdots \cup L^{i_1}.$$

Let B^i be defined formally by

$$B^i f = \sum_{L^{i_1}} \langle f, \rho_\lambda \rangle \rho_\lambda.$$

Then formally

$$A^{i_1} = A^{i_0} + B^{i_0+1} + \cdots + B^{i_1}.$$

Consider first the special ridge function $r^\gamma \equiv \psi^+_{j_0,k_0}(x_1 \cos\theta_0 + x_2 \sin\theta_0)$, with $\gamma = (j_0, k_0, \theta_0)$. For $m = 1, 2, \ldots$.

$$|\langle r^\gamma, \gamma_\lambda \rangle| \leqslant C_m \cdot \delta_{j_0 j} \cdot (\delta_{k_0 k} + \delta_{k_0, 1-k}) \cdot 2^{i/2} \times [(1 + 2^i |\theta_0 - \theta_{i,\ell}|)^{-m} + (1 + 2^i |\theta_0 + \pi - \theta_{i,\ell}|)^{-m}]$$

and as Lemma 4.2 below gives $\|\rho_\lambda\|_{L^\infty(D)} \leqslant C_d \cdot 2^{j/2} 2^{-(i-j)d}$,

$$\sum_{L^{i_1}} |\langle r^\gamma, \rho_\gamma \rangle| \, \|\rho_\lambda\|_{L^\infty(D)} \leqslant 2^{j_0} \cdot 2^{-(i-j_0)(d-1/2)} \cdot C. \tag{4.1}$$

Hence $B^i(r^\gamma)$ is norm-convergent for each γ, and likewise the finite sum $(B^{i_0+1} + \cdots + B^{i_0})(r^\gamma)$.

Now for an arbitrary $r \in \dot{B}^1_{1,1}$, we work by decomposition into special ridge functions. Lemma 3.1 guarantees a decomposition

$$r = \sum_{n=1}^{\infty} c_n \psi^+_{j_n, k_n},$$

with $\sum |c_n| \, 2^{j_n} \leqslant C$, and so taking $r^{(N)}_\theta \equiv \sum_{n=1}^{N} c_n r^{\gamma_n}$ with $\gamma_n = (j_n, k_n, \theta_0)$, we have

$$\| r_{\theta_0} - r^{(N)}_{\theta_0} \|_{L^\infty} \to 0 \qquad N \to \infty.$$

Clearly each sum corresponding to $B^{i_1}(r^{(N)}_{\theta_0})$ is norm-convergent. Moreover, by (4.1),

$$\begin{aligned} \| B^i(r^{(N_2)}_{\theta_0} - r^{(N_1)}_{\theta_0}) \|_{L^\infty} &\leqslant \sum_{n=N_1}^{N_2} \| B^i(c_n r^{\theta_n}) \|_{L^\infty} \\ &\leqslant C \cdot \sum_{n=N_1}^{N_2} 2^{j_n} |c_n|. \end{aligned} \tag{4.2}$$

From $r \in \dot{B}^1_{1,1}$ and (3.1) we get

$$\| B^i(r^{(N_2)}_{\theta_0} - r^{(N_1)}_{\theta_0}) \|_{L^\infty} \leqslant \sum_{N_1}^{\infty} 2^{j_n} |c_n| \; C \leqslant C \, \| r \|_{\dot{B}^1_{1,1}}. \tag{4.3}$$

Hence $(B^i(r^{(N)}_\theta))_N$ is a norm-convergent sequence and the sum corresponding to $B^i(r_\theta)$ is well-defined on $\dot{B}^1_{1,1}$.

It remains to show that A^{i_0} is well-defined. For given $\gamma = (j_0, k_0, \theta_0)$ there are actually only finitely many nonzero terms $\langle r^\gamma, \rho_\lambda \rangle$, $\lambda \in \Lambda^{i_0}$; there are on terms at $j = j_0$, $k \in \{k_0, 1 - k_0\}$ and $0 \leqslant \ell < 2^{i_0 - 1}$. The bound paralleling (4.1) follows for A^{i_0} exactly as for B^i. Inequalities similar to (4.2) and (4.3) go through also. ∎

4.2. *Proof of Theorem* 1.2

Let $r = \sum c_{j,k} \psi^+_{j,k}$ be the representation of the ridge profile r guaranteed by Lemma 3.2. Let $r^{(i_1)} = \sum_{j \leqslant i_1} c_{j,k} \psi^+_{j,k}$ be an approximation to the ridge profile generated using only terms $\psi^+_{j,k}$ at indices $j \leqslant i_1$. Note that the indexlimiting approximation operator A^{i_1} behaves the same on the ridge function r_{θ_0} as on the ridge function with the approximate profile $r^{(i_1)}_{\theta_0}$:

$$A^{i_1} r^{(i_1)}_{\theta_0} = A^{i_1} r_{\theta_0}. \tag{4.4}$$

The triangle inequality gives

$$\|r_{\theta_0} - A^{i_1} r_{\theta_0}\|_{L^\infty(D)} \leqslant \|r_{\theta_0} - r_{\theta_0}^{(i_1)}\|_{L^\infty(D)} + \|r_{\theta_0}^{(i_1)} - A^{i_1} r_{\theta_0}^{(i_1)}\|_{L^\infty(D)}. \tag{4.5}$$

Now as $r \in \dot{B}^s_{1,1}$, (3.1) can be used, along with rapid decay of $w^e_{i,\ell}$, (1.13) and (4.8), giving

$$\begin{aligned}
\|r_{\theta_0} - r_{\theta_0}^{(i_1)}\|_{L^\infty(D)} &\leqslant \sum_{j > i_1} \sum_k |c_{j,k}| \, \|\psi^+_{j,k}\|_{L^\infty[-d,d]} \\
&\leqslant C \sum_{j > i_1} \sum_k |c_{j,k}| \, 2^j \\
&\leqslant C \sum_{j > i_1} 2^{-(s-1)j} \left(\sum_k |c_{j,k}| \, 2^{js} \right) \\
&\leqslant C \cdot 2^{-(s-1) i_1} \sum_{j > i_1} \sum_k |c_{j,k}| \, 2^{js} \\
&\leqslant C' \cdot 2^{-(s-1) i_1} \cdot \|r\|_{\dot{B}^s_{1,1}(\mathbf{R})}.
\end{aligned} \tag{4.6}$$

Meanwhile

$$\|r_{\theta_0}^{(i_1)} - A^{i_1} r_{\theta_0}^{(i_1)}\|_{L^\infty(D)} \leqslant \sum_{i \geqslant i_1} |\langle r_{\theta_0}^{(i_1)}, \rho_\lambda \rangle| \, \|(I - A^{i_1}) \rho_\lambda\|_{L^\infty(D)}. \tag{4.7}$$

Now by Theorem 1.1,

$$\begin{aligned}
|\langle r_{\theta_0}^{(i_1)}, \rho_\lambda \rangle| &\leqslant \sum_\gamma |c_\gamma| \cdot |\langle r^\gamma, \rho_\lambda \rangle| \cdot 1_{\{j \leqslant i_1\}} \\
&\leqslant \sum_{j_0, k_0} |c_{j_0, k_0}| \, (\delta_{jj_0} \cdot (\delta_{kk_0} + \delta_{k(1-k_0)}) \\
&\quad \times (|w_{i,\ell}(\theta_0)| + |w_{i,\ell}(\theta_0 + \pi)|)) \cdot 1_{\{j \leqslant i_1\}} \\
&= (|c_{j,k}| + |c_{j,1-k}|) \cdot (|w_{i,\ell}(\theta_0)| + |w_{i,\ell}(\theta_0 + \pi)|) \cdot 1_{\{j \leqslant i_1\}}.
\end{aligned}$$

Also

$$(I - A^{i_1}) \rho_\lambda = \begin{cases} 0 & j(\lambda) \leqslant i(\lambda) \leqslant i_1 \\ -\rho_\lambda & j(\lambda) \leqslant i_1 < i(\lambda). \end{cases}$$

Lemma 4.2 below gives for $d > 0$

$$\|(I - A^{i_1}) \rho_\lambda\|_{\lambda^\infty(D)} \leqslant \begin{cases} 2^{j/2} 2^{-(i-j)d} \, C_d & j(\lambda) \leqslant i_1 < i(\lambda) \\ 0 & j(\lambda) \leqslant i(\lambda) \leqslant i_1. \end{cases}$$

Hence the right-hand side of (4.7) is upper-bounded by, for $d > s$,

$$
\begin{aligned}
& C \cdot \sum_{j \leqslant i_1} \sum_k |c_{j,k}|\, 2^{j/2} \sum_{i > i_1} 2^{i/2} 2^{-(i-j)d} \\
& \qquad \times \left(\sum_\ell (1 + 2^i\, |\theta_0 - \theta_{i,\ell}|)^{-m} + \sum_\ell (1 + 2^i\, |\theta_0 + \pi - \theta_{i,\ell}|)^{-m} \right) \\
& \leqslant C \sum_{j \leqslant i_1} \sum_k |c_{j,k}|\, 2^{j/2} 2^{-(i_1 - j)d} 2^{i_1/2} \\
& \leqslant C \sum_{j \leqslant i_1} 2^{-(i_1 - j)d} 2^{i_1/2} 2^{j/2} 2^{-js} \sum_k |c_{j,k}|\, 2^{js} \\
& = C 2^{-i_1(d - 1/2)} \sum_{j \leqslant i_1} 2^{j(d + 1/2 - s)} \sum_k |c_{j,k}|\, 2^{js} \\
& \leqslant C 2^{-i_1(d-1/2)} 2^{i_1(d+1/2-s)}\, \|r\|_{\dot{B}^s_{1,1}(\mathbf{R})} \\
& \leqslant C 2^{-i_1(s-1)}\, \|r\|_{\dot{B}^s_{1,1}(\mathbf{R})}.
\end{aligned}
$$

Combining this display with (4.6) and (4.5) gives the desired result (1.8). ∎

LEMMA 4.2. *For* $d > 0$,

$$\|\rho_\lambda\|_{L^\infty(D)} \leqslant C_d \cdot 2^{j/2} 2^{-(i-j)d}, \qquad \lambda \in \Lambda, \quad \varepsilon = 1. \tag{4.8}$$

Proof. From (1.2),

$$\rho_\lambda(x) = \frac{1}{4\pi} \int \psi^+_{j,k}(x_1 \cos\theta + x_2 \sin\theta)\, w^\varepsilon_{i,\ell}(\theta)\, d\theta.$$

Define $f_x(\theta) = \psi^+_{j,k}(x_1 \cos\theta + x_2 \sin\theta)$. Let $t = t(\theta) = x_1 \cos\theta + x_2 \sin\theta$,

$$\left(\frac{\partial}{\partial\theta}\right)^d f_x(\theta) = \sum_{\ell=0}^{d} \left(\frac{\partial}{\partial t}\right)^\ell \psi^+_{j,k}(t)\bigg|_{t = t(\theta)} P_{d,\ell}(t(\theta), ..., t^{(m)}(\theta)),$$

where $P_{d,\ell}$ is a multivariate polynomial of degree $\leqslant d$. Now $t = t(\theta)$ is a smooth function of θ, so $|P_{d,\ell}(t(\theta), ..., t^{(d)}(\theta))| \leqslant C \ \forall \theta \in [0, 2\pi)$. Also, for $m = 1, 2, ...,$ and $\ell = 1, 2, ...$

$$\left|\left(\frac{\partial}{\partial t}\right)^\ell \psi^+_{j,k}(t)\bigg|_{t = t(\theta)} \leqslant 2^{j\ell} 2^j C_{m,\ell} (1 + 2^j\, |t(\theta) - t_{j,k}|)^{-m}.$$

Hence

$$\left\| \left(\frac{\partial}{\partial \theta} \right)^d f_x \right\|_{L^2(d\theta)} \leqslant 2^{jd} 2^{j/2} C.$$

Also, for $\varepsilon = 1$, each $w^1_{i,\ell}$ has a d-fold primitive $(w^1_{i,\ell})^{(-d)}$ obeying the estimate

$$\|(w^1_{i,\ell})^{(-d)}\|_{L^2(d\theta)} \leqslant C 2^{-id},$$

so

$$\begin{aligned} |\rho_\lambda(x)| &= c \left| \int \left[\left(\frac{\partial}{\partial \theta} \right)^d f_x \right] \left[\left(\frac{\partial}{\partial \theta} \right)^{-d} w^1_{i,\ell} \right] d\theta \right| \\ &\leqslant C_d 2^{j/2} 2^{-(i-j)d}, \qquad i > j, \quad d = 1, 2, \ldots. \end{aligned}$$

5. NONLINEAR APPROXIMATION OF RIDGE FUNCTIONS

In this section we prove Theorem 1.3.

5.1. *Preliminaries*

A simple adaptation of the earlier Lemma 3.2 concerning $\dot{B}^1_{1,1}$ will prove

LEMMA 5.1. *Let $r \in \dot{B}^s_{p,p}(\mathbf{R})$, where $s = 1/p$ and $0 < p < 1$, and vanish at $\pm\infty$. Then*

$$r(t) = \sum_{j,k} c_{j,k} \psi^+_{j,k},$$

where the sum is convergent in L^∞, and for $C < \infty$,

$$\sum_{j,k} |c_{j,k}|^p \, 2^{jp} \leqslant C^p. \tag{5.1}$$

Indeed, the argument is the same as the argument for Lemma 3.2, only substituting p-summability of the kernel $A(j', k'; j, k)$ for the 1-summability used in the proof of Lemma 3.2. That is, proceeding exactly as in that proof until (6.4), one invokes at that point Lemma 6.2 and (6.5) and then one obtains from this, and the p-triangle inequality, the desired relation (5.1).

We need also a fact of the type well-known in interpolation theory—see [2, 17].

LEMMA 5.2. *Suppose* $(v_i: i = 1, 2, ...)$ *is a sequence of nonincreasing positive numbers with, for fixed* $0 < p < 1$, $\sum v_i^p \leqslant V^p$. *Let* $S_N = \sum_{i=N}^{\infty} v_i$. *Then for* $C_p > 0$,

$$S_N \leqslant C_p \cdot V \cdot N^{1-1/p}, \qquad N = 1, 2, 3, ... \tag{5.2}$$

Proof. It follows from the assumption that $v_1 \geqslant \cdots \geqslant v_i$ that $i \cdot v_i^p \leqslant V^p$, for $i = 1, 2, 3, ...$ Hence $S_n \leqslant V \cdot \sum_{i=N}^{\infty} i^{-1/p} \leqslant C_p \cdot V \cdot N^{-(1/p-1)}$. ∎

5.2. *Proof of Theorem* 1.3

Let as before $r^\gamma(x) = \psi^+_{j_0, k_0}(x_1 \cos\theta_0 + x_2 \sin\theta_0)$ be a special ridge function, with parameter $\gamma = (j_0, k_0, \theta_0)$. By hypothesis,

$$r_{\theta_0} = \sum_{j_0, k_0} c_\gamma r^\gamma, \tag{5.3}$$

where, as $r \in \dot{B}^s_{p,p}(\mathbf{R})$ for $s = 1/p$,

$$\sum_{j_0, k_0} |c_\gamma|^p \, 2^{j_0 p} \leqslant C^p \tag{5.4}$$

for some $C > 0$.

Define now the *ridge-molecule* $x^\gamma = (x^\gamma_\lambda: \lambda \in \Lambda)$ by

$$x^\gamma_\lambda = \langle r^\gamma, \rho_\lambda \rangle \cdot \|\rho_\lambda\|_{L^\infty(D)}, \qquad \lambda \in \Lambda; \tag{5.5}$$

(x^γ_λ) is the sequence of ridgelet coefficients of the special ridge function r^γ, normalized by their "observable effects" in the disk D. Lemma 5.3 below shows that for $0 < p < 1$,

$$\|x^\gamma\|_{\ell^p} \leqslant C_p \cdot 2^{j_0}, \tag{5.6}$$

where C_p does not depend on γ; this justifies the appellation "molecular". Now the ℓ^p-quasi norm obeys the p-triangle inequality

$$\|x + y\|^p_{\ell^p} \leqslant \|x\|^p_{\ell^p} + \|y\|^p_{\ell^p} \tag{5.7}$$

for arbitrary sequences x, y. Define the array

$$x = (\langle r_{\theta_0}, \rho_\lambda \rangle \cdot \|\rho_\lambda\|_{L^\infty(D)}: \lambda \in \Lambda);$$

this is the sequence of effect-normalized coefficients of the object r_{θ_0}. The array x obeys, by (5.3)–(5.4), the molecular decomposition

$$x = \sum_{j_0, k_0} c_\gamma x^\gamma. \tag{5.8}$$

Combining this with (5.6) and (5.7),

$$\begin{aligned}\|x\|_{\ell^p}^p &= \left\| \sum_{j_0,\, k_0} c_\gamma x^\gamma \right\|_{\ell^p}^p \\ &\leqslant \sum_{j_0,\, k_0} |c_\gamma|^p \, \|x^\gamma\|_{\ell^p}^p \\ &\leqslant C \sum_{j_0,\, k_0} |c_\gamma|^p \, 2^{j_0 p} \\ &\leqslant C' \cdot \|r\|_{\dot{B}^s_{p,\,p}(\mathbf{R})}^p. \end{aligned} \tag{5.9}$$

Now consider

$$\tilde{r}^{(\delta)} = \sum_\lambda \eta_\delta(\langle r_{\theta_0}, \rho_\lambda \rangle, \, \|\rho_\lambda\|_{L^\infty(D)}) \, \rho_\lambda.$$

We have, with $\Lambda_\delta^c = \{\lambda : \langle r_{\theta_0}, \rho_\lambda \rangle | \cdot \|\rho_\lambda\|_{L^\infty(D)} < \delta\}$,

$$\begin{aligned}\|r_{\theta_0} - \tilde{r}^{(\delta)}\|_{L^\infty(D)} &= \left\| \sum_{\Lambda_\delta^c} \langle r_{\theta_0}, \rho_\lambda \rangle \, \rho_\lambda \right\|_{L^\infty(D)} \\ &\leqslant \sum_{\Lambda_\delta^c} |\langle r_{\theta_0}, \rho_\lambda \rangle| \cdot \|\rho_\lambda\|_{L^\infty(D)}. \end{aligned} \tag{5.10}$$

Let $v_1, v_2, \ldots$ be an enumeration of the entries $|\langle r_{\theta_0}, \rho_\lambda \rangle| \cdot \|\rho_\lambda\|_{L^\infty(D)}$ in nonincreasing order. There is an $N = N(\delta)$ so that the sum on the right of (5.10) is of the form $S_N = \sum_{i=N}^\infty v_i$ discussed in the lead-up to (5.2). Therefore, with $V = (\sum v_i^p)^{1/p}$ bounded by $C \cdot \|r\|_{\dot{B}^s_{p,\,p}(\mathbf{R})}$ according to (5.9), we have by (5.2)

$$\|r_{\theta_0} - \tilde{r}^{(\delta)}\|_{L^\infty(D)} \leqslant C \cdot N^{1-1/p} = C \cdot N^{-(s-1)}, \qquad N = N(\delta), \quad \forall \delta > 0.$$

This completes the proof of Theorem 1.3.

5.3. *Proof of Corollary* 1.5

We now consider Corollary 1.5, which considered nonlinear approximations to a function f which is a superposition of ridge functions. Define the array

$$x = (\langle f, \rho_\lambda \rangle \cdot \|\rho_\lambda\|_{L^\infty(D)} : \lambda \in \Lambda);$$

this is the sequence of effect-normalized coefficients of the object f. By assumption, f is a superposition of ridge functions $f = \sum_i r_{i,\,\theta_i}$, and each ridge function $r_{i,\,\theta_i}$ obeys a decomposition exactly like (5.3)

$$r_{\theta_i} = \sum_{j_0,\, k_0} c_{i,\,\gamma} r^\gamma, \tag{5.11}$$

where in this sum, γ runs through all triples (j_0, k_0, θ_i) with θ_i fixed. It follows that, with x^γ as in (5.5),

$$x = \sum_{j_0,\, k_0,\, i} c_{i,\gamma} x^\gamma. \tag{5.12}$$

Following exactly the logic of (5.8)–(5.9), we have

$$\|x\|_{\ell^p}^p \leqslant C \cdot \sum_i \sum_{j_0,\, k_0} |c_{i,\gamma}|^p \, 2^{j_0 p}.$$

But now observe that the right-hand side is equivalent to $C \cdot \sum_i \|r_i\|_{\dot{B}^s_{p,p}}^p \equiv C \cdot R$. From that point on, argue as in the completion of the proof of Theorem 1.3, starting from (5.9) on.

5.4. *Ridgelet Analysis of a Special Ridge Function*

LEMMA 5.3. *For the special ridge function r^γ, the corresponding sequence x^γ of ridgelet coefficients normalized by $\|\rho_\lambda\|_{L^\infty(D)}$ obeys*

$$\|x^\gamma\|_{\ell^p} \leqslant C_p \cdot 2^{j_0}, \tag{5.13}$$

where C_p is independent of (j_0, k_0, θ_0).

Proof. Now $x^\gamma_\lambda = a^\gamma_\lambda \cdot \|\rho_\lambda\|_{L^\infty(D)}$, where a^γ_λ is as in (1.3). The a^γ_λ obey, by rapid decay of $w^e_{i,\ell}$, an estimate of the form

$$|a^\gamma_\lambda| \leqslant C_m \cdot 2^{i/2} \cdot ((1 + 2^i\,|\theta_0 - \theta_{i,\ell}|)^{-m} + (1 + 2^i\,|\theta_0 + \pi - \theta_{i,\ell}|)^{-m}), \quad \forall i, \ell \tag{5.14}$$

for each $m > 1$. Moreover, by Lemma 4.2,

$$\|\rho_\lambda\|_{L^\infty(D)} \leqslant C_d \cdot 2^{j_0/2} \cdot 2^{-(i - j_0)\,d}, \tag{5.15}$$

for each $d > 0$. Set now, for fixed m,

$$u^i_\ell(t) = (1 + 2^i\,|t - \theta_{i,\ell}|)^{-m},$$

and note that by (5.14), $|a^\gamma_\lambda| \leqslant C_m \cdot 2^{i/2} \cdot (u^i_\ell(\theta_0) + u^i_\ell(\theta_0 + \pi))$. Now for $mp > 1$, and $t \in [0, 2\pi)$,

$$\|(u^i_\ell(t))_\ell\|_{L^p} \leqslant C_{m,\,p} < \infty. \tag{5.16}$$

The nonzero entries in the sequence x^γ correspond to indices λ where $j(\lambda) = j_0$, $k(\lambda) = k_0$, and $h = i - j \geqslant 0$, $\ell = 0, \ldots, 2^{j+h-1} - 1$. Define, for each $h > 0$, the array v^h structured as

$$v^h = (v^h_\ell : \ell = 0, \ldots, 2^{j_0 + h - 1} - 1),$$

and note that the nonzero entries in x^γ can be arranged in a concatenation of several such arrays

$$v = (v^0, v^1, v^2, \ldots),$$

where v^h contains the nonzero entries x^γ_λ such that $i = j_0 + h$; $k = k_0$, $\ell = 0, \ldots, 2^{j_0+h-1} - 1$. Now combining (5.14) and (5.15), we have, with $A_h = C_d \cdot 2^{j_0/2} \cdot 2^{-hd}$,

$$\begin{aligned} \|v^h\|^p_{\ell^p} &\leqslant A_h^p \cdot (\|u^{j_0+h}(\theta_0)\|^p_{\ell^p} + \|u^{j_0+h}(\theta_0+\pi)\|^p_{\ell^p}) \cdot 2^{(j_0+h)\,p/2} \\ &\leqslant C_{m,\,p} A_h^p 2^{(j_0+h)\,p/2}. \end{aligned}$$

We also remark that if an array v is the concatenation of two subarrays, $v = (v^0, v^1)$, then

$$\|v\|^p_{\ell^p} = \|v^0\|^p_{\ell^p} + \|v^1\|^p_{\ell^p}. \tag{5.17}$$

Using this and picking $d > 1/2$, so that $\sum_{h \geqslant 0} (2^{-hd})^p\, 2^{hp/2} \leqslant C'_{d,\,p}$, we have

$$\begin{aligned} \|(v^0, v^1, \ldots)\|^p_{\ell^p} &= \sum_{h \geqslant 0} \|v^h\|^p_{\ell^p} \\ &\leqslant C^p_{m,\,p} \sum_{h \geqslant 0} A_h^p \cdot 2^{(j_0+h)\,p/2} \\ &\leqslant C' 2^{j_0 p}, \end{aligned}$$

and so (5.13) follows. ∎

6. APPENDIX

Proof of Lemma 3.1. This follows from Lemma 3.2. Write

$$\left\| \sum c_{j,k} \psi^+_{j,k} \right\|_{L^\infty} \leqslant \sum |c_{j,k}|\, \|\psi^+_{j,k}\|_{L^\infty} \leqslant C \sum |c_{j,k}|\, 2^j < \infty. \tag{6.1}$$

Let a_n be the nth largest among the entries $|c_{j,k}| \|\psi^+_{j,k}\|_{L^\infty}$ and suppose it occurs at index (j_n, k_n). Set

$$r_N \equiv \sum_{n=1}^{N} c_{j_n, k_n} \psi^+_{j_n, k_n};$$

then we have

$$\|r - r_N\|_\infty \leqslant \sum_{N+1}^{\infty} a_n.$$

By (6.1), $r \in \dot{B}^1_{1,1}(\mathbf{R})$ implies $\sum_1^\infty a_n < \infty$, so $\sum_{N+1}^\infty a_n \to 0$ as $N \to \infty$.

Proof of Lemma 3.2. We have, from $\Delta^\pm \Delta^\mp = \mathrm{Id}$ on $\mathrm{Dom}(\Delta^+) \cap \mathrm{Dom}(\Delta^-)$, from self-adjointness of $\Delta^\pm$, and from $\psi_{j,k} \in \mathrm{Dom}(\Delta^+) \cap \mathrm{Dom}(\Delta^-)$,

$$\begin{aligned}\langle \psi^+_{j',k'}, \psi^-_{j,k} \rangle &= \langle \Delta^+ \psi_{j',k'}, \Delta^- \psi_{j,k} \rangle \\ &= \langle \Delta^- \Delta^+ \psi_{j',k'}, \psi_{j,k} \rangle \\ &= \langle \psi_{j',k'}, \psi_{j,k} \rangle = \delta_{j'j} \cdot \delta_{k'k}.\end{aligned}$$

This gives the identity

$$f = \sum_{j,k} \langle f, \psi^-_{j,k} \rangle \, \psi^+_{j,k}$$

valid for any $f = \sum_{n=1}^N c_n \psi^+_{j_n,k_n}$ which is a finite sum of $\psi^+_{j,k}$'s.

We claim that this identity is also valid for $f = \psi_{j',k'}$, which is an infinite sum:

$$\psi_{j',k'} = \sum_{j=j'-1}^{j'+1} \sum_k \langle \psi_{j',k'}, \psi^-_{j,k} \rangle \, \psi^+_{j,k}. \tag{6.2}$$

To justify this we argue as follows. Let $v(\omega)$ be a smooth window, equal to 1 on $|\omega| \in [\pi/(3 \cdot 128), 128\pi/3]$ and vanishing outside $\pi/(3 \cdot 256)$ to $256\pi/3$. Define Δ^+_j and Δ^-_j by

$$(\Delta^\pm_j f)(t) = \int \hat{f}(\omega) \, e^{i\omega t} \, |\omega|^{\pm 1/2} \, v(2^{-j}\omega) \, d\omega.$$

Then each $\Delta^\pm_j$ is a bounded convolution operator, and agrees perfectly with $\Delta^\pm$ on the three adjacent levels $j' \in \{j-1, j, j+1\}$:

$$\Delta^\pm_j \psi_{j',k'} = \Delta^\pm \psi_{j',k'} \qquad j' \in \{j-1, j, j+1\}, \quad k \in \mathbf{Z}.$$

Also $\Delta^\pm_j \Delta^\mp_j = \mathrm{Id}$ on $\mathrm{span}\{\psi_{j',k} : j' = j-1, j, j+1, k \in \mathbf{Z}\}$. According to Lemma 6.1 and translation invariance $\psi_{j,k+1}(t) = \psi_{j,k}(t - 2^{-j})$,

$$\Delta^-_{j'} \psi_{j',k'} = \sum_{j=j'-1}^{j=j'+1} \sum_k \langle \Delta^-_{j'} \psi_{j',k'}, \psi_{j,k} \rangle \, \psi_{j,k}.$$

It follows from self-adjointness of Δ_j^-,

$$\begin{aligned}\langle \psi^-_{j',k'}, \psi_{j,k}\rangle &= \langle \Delta^-_{j'}\psi_{j'}, \psi_{j,k}\rangle \\ &= \langle \psi_{j',k'}, \Delta^-_{j'}\psi_{j,k}\rangle \\ &= \langle \psi_{j',k'}, \psi^-_{j,k}\rangle, \qquad j\in\{j'-1, j', j'+1\}.\end{aligned}$$

So we may write

$$\psi^-_{j',k'} = \sum_{j=j'-1}^{j=j'+1}\sum_k \langle \psi_{j',k'}, \psi^-_{j,k}\rangle\, \psi_{j,k}.$$

Operating on both sides by $\Delta^+_{j'}$ and justifying a term-by-term treatment using Lemma 6.2,

$$\Delta^+_{j'}\psi^-_{j',k'} = \sum_{j=j'-1}^{j'+1}\sum_k \langle \psi_{j',k'}, \psi^-_{j,k}\rangle\, \Delta^+_{j'}\psi_{j,k}$$

which, as $\Delta^+_{j'}\psi^-_{j',k'} = \psi_{j',k'}$ and $\Delta^+_{j'}\psi_{j,k'} = \psi^+_{j,k}$ in the indicated range, this justifies (6.2).

Define now

$$A(j',k';j,k) = \langle \psi_{j',k'}, \psi^-_{j,k}\rangle.$$

Note that $\psi^-_{j,k}(t) = \psi^-_{0,0}(2^j t - k)$; hence

$$\begin{aligned}A(j',k';j,k) &= \int \psi_{j',k'}(t)\,\psi^-_{j,k}(t)\,dt \\ &= 2^{j'/2}\int \psi_{0,k'}(2^{j'}t)\,\psi^-_{j-j',k}(2^{j'}t)\,2^{j'}\,dt \\ &= 2^{-j'/2}\langle \psi_{0,k'}, \psi^-_{j-j',k}\rangle \\ &= 2^{-j'/2}A(0,k';j-j',k). \qquad (6.3)\end{aligned}$$

Now (6.2) says that

$$\psi_{j',k'} = \sum A(j',k';j,k)\,\psi^+_{j,k},$$

and we may use A to convert an expansion in wavelets $\psi_{j,k}$ into an expansion in unnormalized vaguelettes $\psi^+_{j,k}$. Applying this idea,

$$r(t) = \sum_{j', k'} \alpha'_{j', k'} \psi_{j', k'}$$

$$= \sum_{j', k'} \alpha_{j', k'} \sum_{j, k} A(j'k'; j, k)\, \psi^+_{j, k}$$

$$= \sum_{j, k} \psi^+_{j, k} \sum_{j', k'} A(j'k'; j, k)\, \alpha_{j', k'}.$$

Hence,

$$r(t) = \sum a_{jk} \psi^+_{j, k},$$

with, by (6.3),

$$a_{j, k} = \sum_{j', k'} A(j'k'; j, k)\, \alpha_{j'k'}$$

$$= \sum_{j' = j-1}^{j+1} \sum_{k'} 2^{-j'/2} A(0, k'; j - j', k)\, \alpha_{j'k'}.$$

Lemma 6.2 gives

$$\sum_{k'} |A(0, k', h, k)| \leqslant C \qquad \forall k, \quad \forall h = -1, 0, 1, \tag{6.4}$$

and we have

$$\sum_{k} |a_{j, k}| \leqslant C \sum_{j' = j-1}^{j+1} \sum_{k'} |\alpha_{j'k'}|\, 2^{-j'/2}.$$

In short, for $s \geqslant 1$,

$$\sum_{j, k} 2^{js} |a_{j, k}| \leqslant C \sum_{j', k'} |\alpha_{j', k'}|\, 2^{j(s-1/2)}. \quad \blacksquare$$

LEMMA 6.1. *Let T be an L^2-bounded convolution operator. Let $(\psi_{j, k})$ be a system of Meyer wavelets.*

$$T\psi_{0, 0} = \sum_{j = -1}^{+1} \sum_{k} \langle T\psi_{0, 0}, \psi_{j, k} \rangle\, \psi_{j, k}.$$

Proof. As $T\psi_{0, 0}$ is in L^2, it automatically has a representation $T\psi_{0, 0} = \sum_j \sum_k \langle T\psi_{0, 0}, \psi_{j, k} \rangle\, \psi_{j, k}$ with potentially infinitely many scales. However, as T is a convolution operator we may also write $\widehat{T\psi}(\omega) = \tau(\omega)\, \hat{\psi}(\omega)$ with

$\tau(\omega)$ an essentially bounded function, and so the support of $\widehat{T\psi}(\omega)$ is contained in the support of $\psi(\omega)$. Thus $\widehat{T\psi}_{0,0}(\omega)$ is supported in $|\omega| \in [2/3\pi, 8/3\pi]$. Hence it is orthogonal to every $\psi_{j,k}$ whose support in the frequency domain does not intersect the interior of this set. In short it is orthogonal to every $\psi_{j,k}$ with $j < -1$ or $j > 1$. ∎

LEMMA 6.2. *For* $0 < p \leqslant 1$,

$$\sum_{k'} |A(0, k', h, k)|^p \leqslant C_p \qquad \forall k, \forall h = -1, 0, 1, \tag{6.5}$$

Proof. Writing ψ for $\psi_{0,0}$ and passing to the frequency domain,

$$\begin{aligned} A(0, k', h, k) &= \frac{1}{2\pi} \int \hat{\psi}(\omega)\, e^{-i\omega k'} \hat{\psi}(\omega/2^h)^*\, e^{+i\omega k/2^h} 2^{-h/2}\, d\omega \\ &= \frac{1}{2\pi} \int \hat{\psi}(\omega)\, \hat{\psi}(\omega/2^h)^*\, 2^{-h/2} e^{-i\omega(k' - k/2^h)}\, d\omega \\ &= \Psi_h(k/2^h - k'), \qquad \text{say,} \end{aligned}$$

where

$$\hat{\Psi}_h(\omega) \equiv \hat{\psi}(\omega)\, \hat{\psi}(\omega/2^h)^*\, 2^{-h/2}.$$

Now $\hat{\Psi}_h$ is C^∞ and of compact support. Hence Ψ_h is of rapid decay, and so for each $m > 0$ we have C_m with

$$|\Psi_h(t)| \leqslant C_m \cdot (1 + |t|)^{-m} \qquad \forall t \in \mathbf{R}.$$

Now picking $mp > 1$,

$$\begin{aligned} \sum_{k'} |A(0, k', h, k)|^p &= \sum_{k'} |\Psi_h(k/2^h - k')|^p \\ &\leqslant C_m^p \sum_{k'} (1 + |k/2^h - k'|)^{-mp} \leqslant C_{m,p} < \infty. \end{aligned}$$ ∎

ACKNOWLEDGMENTS

This research was supported by National Science Foundation Grant DMS 95-05151 and by AFOSR MURI 95-P49620-96-1-0028.

REFERENCES

1. A. Barron, Universal approximation bounds for superpositions of a sigmoidal function, *IEEE Trans. Infor. Theory* **39** (1993), 930–945.
2. J. Bergh and J. Löfström, "Interpolation Spaces. An Introduction," Grundlehren der Mathematischen Wissenschaften, No. 223, Springer-Verlag, Berlin/New York, 1976.
3. E. J. Candès, Harmonic analysis of neural networks, *Appl. Comput. Harmon. Anal.* **6** (1999), 197–218.
4. E. Candès, "Ridgelets: Theory and Applications," Ph.D. thesis, Department of Statistics, Stanford University, 1998.
5. E. Candès and D. Donoho, Ridgelets: the key to high-dimensional intermittency?, *Phil. Trans. Roy. Soc. Lond. A* **357** (1999), 2495–2509.
6. C. K. Chui, "Wavelets: A Mathematical Tool for Signal Analysis," SIAM, Philadelphia, 1997.
7. R. A. DeVore, Degree of nonlinear approximation, *in* "Approx. Theory VI" (C. K. Chui, L. L. Schumaker, and J. D. Ward, Eds.), Vol. I, pp. 175–201, Academic Press, San Diego, 1988.
8. R. A. DeVore and G. G. Lorentz, "Constructive Approximation," Springer-Verlag, Berlin/New York, 1993.
9. D. L. Donoho, Orthonormal ridgelets and linear singularities, *SIAM J. Math. Anal.* **31**(5) (2000), 1030–1061.
10. D. L. Donoho, Tight frames of k-plane ridgelets and the problem of representing d-dimensional singularities in $\mathbf{R}^n$, *Proc. Nat. Acad. Sci. USA* **96** (1999), 1828–1833.
11. J. H. Friedman and W. Stuetzle, Projection pursuit regression, *J. Amer. Statist. Assoc.* **76** (1981), 817–823.
12. L. K. Jones, A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training, *Ann. Statist.* **20** (1992), 608–613.
13. B. F. Logan and L. A. Shepp, Optimal reconstruction of a function from its projections, *Duke Math. J.* **42** (1975), 645–659.
14. P. G. Lemarié and Y. Meyer, Ondelettes et bases Hilbertiennes, *Rev. Mat. Iberoamer.* **2** (1986), 1–18.
15. Y. Meyer, "Ondelettes et Opérateurs," Hermann, Paris, 1990. English Translation: "Wavelets and Operators," Cambridge University Press, Cambridge, UK, 1992.
16. N. Murata, An integral representation of functions using three-layered networks and their approximation bounds, *Neural Networks* **9** (1996), 947–956.
17. J. Peetre and G. Sparr, Interpolation of normed Abelian groups, *Ann. Mat. Pura Appl. IV* **92** (1972), 217–262.