RIDGE FUNCTIONS AND
ORTHONORMAL RIDGELETS

by

DAVID L. DONOHO
Department of Statistics
Stanford University

December 1998

This research was supported in part by
National Science Foundation grant DMS-9505151
and Air Force Office of Scientific Research grant
MURI95-F49620-96-1-0028

Department of Statistics
Sequoia Hall
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://www-stat.stanford.edu
Ridge Functions and Orthonormal Ridgelets

DAVID L. DONoho

Department of Statistics
Stanford University
Stanford, California 94305–4065

Orthonormal ridgelets are a specialized set of angularly-integrated ridge functions which make up an orthonormal basis for $L^2(\mathbb{R}^2)$. 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 giving the ridgelet coefficients of a ridge function in terms of the 1-D wavelet coefficients of the ridge profile $r(t)$, and we study the properties of the linear approximation operator which 'kills' coefficients at high angular scale or high ridge scale. We also show that partial orthonormal ridgelet expansions can give efficient nonlinear approximations to pure ridge functions. In effect, 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 $r(t)$. This shows that simple thresholding in the ridgelet basis is, for certain purposes, equally as good as ideal nonlinear ridge approximation.


**Acknowledgments.** This research was supported by National Science Foundation grant DMS 95–05151 and by AFOSR MURI 95–P49620–96–1–0028.
1 Introduction

In [5], we introduced orthonormal ridgelets, defined as follows. Let \( (\psi_{j,k}(t) : j \in \mathbb{Z}, k \in \mathbb{Z}) \) be an orthonormal basis of Meyer wavelets for \( L^2(\mathbb{R}) \), and let \( (w_{i_0,\ell}^0(\theta) : \ell = 0, \ldots, 2^{i_0} - 1; \ w_{i_0,\ell}^1(\theta), \ i \geq i_0, \ \ell = 0, \ldots, 2^i - 1) \) be an orthonormal basis for \( L^2[0, 2\pi] \) made of periodized Lemarié scaling functions \( w_{i_0,\ell}^0 \) at level \( i_0 \) and periodized Meyer wavelets \( w_{i,\ell}^1 \) at levels \( i \geq i_0 \). (We suppose a particular normalization of these functions; see Section 2.1 below). Let \( \hat{\psi}_{j,k}(\omega) \) denote the Fourier transform of \( \psi_{j,k}(t) \), and define ridgelets \( \rho_\lambda(x), \ \lambda = (j,k,i,\ell,\epsilon) \) as functions of \( x \in \mathbb{R}^2 \) using the frequency-domain definition

\[
\rho_\lambda(\xi) = |\xi|^{-\frac{1}{2}} (\hat{\psi}_{j,k}(\xi) w_{i,\ell}^\epsilon(\theta) + \hat{\psi}_{j,k}(-|\xi|) w_{i,\ell}^\epsilon(\theta + \pi))/2.
\]

(1.1)

Here the indices run as follows: \( j, k \in \mathbb{Z}, \ell = 0, \ldots, 2^{i-1} - 1, i \geq i_0 \); and, if \( i > i_0 \), \( i \geq j \). Also, if \( i > i_0 \) and \( i > j \), then necessarily \( \epsilon = 1 \). Let \( \Lambda \) denote the set of all such indices \( \lambda \). (To avoid confusion, note that in [5] two orthonormal systems were discussed; the one we study here was defined and studied only in Section 7 of [5]).

In that article, it was shown that this collection of functions makes an orthonormal set for \( L^2(\mathbb{R}^2) \).

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

\[
\rho_\lambda(x) = \frac{1}{4\pi} \int_0^{2\pi} \psi_{j,k}^+(x_1 \cos \theta + x_2 \sin \theta) w_{i,\ell}^\epsilon(\theta) d\theta.
\]

(1.2)

Here, each \( \psi_{j,k}^+(x_1 \cos \theta + x_2 \sin \theta) \) is a ridge function of \( x \in \mathbb{R}^2 \), i.e., a function of the form \( r(x_1 \cos \theta + x_2 \sin \theta) \). Therefore \( \rho_\lambda \) is obtained by “averaging” ridge functions with ridge angles \( \theta \) 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 [7] 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 [6]. A celebrated result of Barron established the ability of an appropriately-chosen superposition of ridge functions \( f_N(x) = \sum_{n=1}^N a_n \sigma(u_n x - b_n) \) 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 [2] 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 known to be stable functionals of the underlying approximand \( f \); 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 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 [2] and Murata [10] 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 \( a > 0, b \in \mathbb{R} \), and \( u \in S^{d-1} \), and taking values \( R_f(a,b,u) = \langle f, \psi_{a,b,u} \rangle \) (we are following notation of [2]). This continuous transform has a reconstruction formula \( f(x) = \int R_f(a,b,\theta) \psi_{a,b,\theta}(x) d\mu(a,b,\theta) \), 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 \( \langle \psi_{u_n,b_n,u_n}(x) \rangle \) 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 $L^2(\mathbb{R}^2)$, 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 [5] 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 $\times$ angular frequency domain. Under this broader conception, ridgelets no longer are of the form $\psi_{a,b}(x)$, 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 plays 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 [2, 3]) 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 \( r^\gamma(x) = \psi^+_{j_0, k_0}(z_1 \cos \theta_0 + z_2 \sin \theta_0) \). Define, for \( \lambda = (j, k; i, \ell, \varepsilon) \), the array of coefficients

\[
a^\gamma_\lambda = \delta_{j_0, j} \delta_{k_0, k} \cdot w^-_{i, \ell}(\theta_0) + \delta_{j_0, 1} \delta_{k_0, 1-k} \cdot w^+_{i, \ell}(\theta_0 + \pi),
\]

where the \( \delta \)'s denote Kronecker symbols. We will show below that the special ridge function \( r^\gamma \) has a representation by ridgelets \( \rho_\lambda \) using precisely the coefficients \( a^\gamma_\lambda \):

\[
r^\gamma = \sum_\lambda a^\gamma_\lambda \rho_\lambda.
\]

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 \( \ell \) index, all the action occurs ‘near’ the index \( \ell_0 \) with \( \theta_{i, \ell} \) closest to \( \theta_0 \), or at its counterpart closest to \( \theta_0 + \pi \). In a certain sense, the pure ridge function \( r^\gamma \) is a superposition of a small number of ridgelets.

An apparent exception to our sparsity assertion occurs in the index \( i \), where the coefficients \( a^\gamma_\lambda \) are increasing exponentially in \( i \) (as \( u^-_{i, \ell_0}(\theta_0) \approx 2^{i/2} \)). 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 \( \rho_\lambda \) from the origin. Hence, if we restrict attention to a disk \( D = \{(x_1, x_2) : x^2_1 + x^2_2 < d^2\} \) then, in a certain sense most of the action in \( \rho_\lambda \) 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 an orthonormal ridgelet 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 \( \dot{B}^s_{p,q}(\mathbb{R}) \) if it can be represented as a wavelet series

\[
r = \sum_{j,k} a_{j,k} \psi_{j,k}
\]
where \((\psi_{j,k})\) is a family of Meyer orthonormal wavelets, and where the coefficients \((\alpha_{j,k})\) obey
\[
\left( \sum_{j,k} |\alpha_{j,k}|^p \omega^{(s+1/2-1/p)} \right)^{1/p} < \infty.
\] (1.6)

We will only be interested in the cases where \(p \leq 1\) and \(s \geq 1/p\). 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 \(r = \sum c_t g_{a_t,b_t}(t)\) and \(\sum |c_t| < \infty.\) 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 \(\dot{B}_{p,p}^s\) and vanishes at \(\pm \infty.\)

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

Define the fractionally-integrated Meyer wavelet
\[
\psi_{j,k}(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\omega|^{-\frac{1}{2}} \hat{\psi}_{j,k}(\omega) e^{i\omega t} d\omega.
\]

**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 \(\dot{B}_{1,1}^s\)). 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 \(\theta.\) Then
\[
\langle r_{\theta}, \rho_{\lambda} \rangle = \left( \langle r_{\theta}, \psi_{j,0}^+ \rangle_{\mathcal{W}_{s,p}}(\theta) + \langle r, \psi_{j,1-k}^- \rangle_{\mathcal{W}_{s,p}}(\theta + \pi) \right)/2.
\] (1.7)

Here the scalar product on the left of (1.7),
\[
\langle r_{\theta}, \rho_{\lambda} \rangle \equiv \int_{\mathbb{R}^2} r_{\theta}(x) \rho_{\lambda}(x) dx,
\]
may be interpreted as a pairing between \(L^\infty(\mathbb{R}^2)\) and \(L^1(\mathbb{R}^2)\), while the scalar product on the right, \(\langle r, \psi_{j,k}^- \rangle = \int_{\mathbb{R}} r(t) \psi_{j,k}^-(t) dt,\) is a pairing between \(L^\infty(\mathbb{R})\) and \(L^1(\mathbb{R})\). 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 \( A^i f = \sum_{i,j \leq t_i} (f, \rho_{j}) \rho_{i} \) be the partial sum approximation to \( f \) by ridgelets with both \( i \) and \( j \) index-limited so that only scales larger than \( 2^{-i} \) are used.

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

\[
\| r_{\theta} - A^{i_1} r_{\theta} \|_{L^\infty(D)} \leq C 2^{-i_1(s-1)} \| r \|_{B^s_{1,1}(\mathbb{R})}, \quad i_1 = i_0, i_0 + 1, \ldots.
\]  

(1.8)

Here \( L^\infty(D) \) means supremum norm over a fixed disk, \( D = \{ x^2_1 + x^2_2 < d^2 \} \).

This shows that linear approximation by ridgelets converges rapidly to nice ridge functions. In fact it says something surprising. Note that if we knew that we were dealing with a ridge function, and we knew the ridge direction \( \theta \), we could form a kind of ideal linear approximation to a ridge function \( r_{\theta} \), by the following two-step procedure. First, we would approximate the ridge profile \( r \), letting \( r^{(i_1)} = \sum_{j \leq t_i} \alpha_{j,k} \psi_{j,k} \) be the approximate profile gotten by chopping off the terms beyond level \( j = i_1 \) in the representation of the ridge profile by (say) 1-d wavelets \( \psi_{j,k} \). Second, we would build a ridge function using this approximate profile, getting the approximant \( r^{(i_1)}_{\theta} = r^{(i_1)}(x_1 \cos(\theta) + x_2 \sin(\theta)) \). Of course, this is an ideal procedure only; it is not realizable in practice, because it depends on knowledge of \( \theta \), which one does not ordinarily have. However, it is interesting for making a performance comparison.

This procedure behaves as follows. The error of 1-dimensional approximation on the interval \([-d,d]\) caused by ignoring the fine scales would behave as

\[
\| r - r^{(i_1)} \|_{L^\infty[-d,d]} \leq C 2^{-i_1(s-1)} \| r \|_{B^s_{1,1}(\mathbb{R})}, \quad i_1 = 1, 2, 3, \ldots,
\]

and, from standard ideas in approximation theory, we know that in general no estimate substantially better than this can be obtained using any linear procedure. Now when approximating ridge functions by ridge functions, the \( L^\infty \)-error of 2-dimensional approximation is isometric to the \( L^\infty \)-error of 1-dimensional approximation:

\[
\| r - r^{(i_1)} \|_{L^\infty[-d,d]} = \| r_{\theta} - r^{(i_1)}_{\theta} \|_{L^\infty(D)}.
\]  

(1.9)
It follows that for the error of ideal linear approximation 'knowing $\theta$' we have
\[
\|r_\theta - r_{\hat{\theta}}^{(i_1)}\|_{L^\infty(D)} \leq C \cdot 2^{-i_1(s-1)} \|r\|_{\hat{B}^s_{1,1}(\mathbb{R})}, \quad i_1 = 1, 2, 3, \ldots,
\]
and no substantially better estimate is possible for such ideal linear procedures. However, comparing with (1.8) we see that this estimate is not really better than the type of estimate available from linear approximation using ridgelets. The point is that the ridgelet approximation is realizable - to construct it requires no knowledge of $\theta$, nor even that the approximand be a ridge function at all - yet it attains the same quality of approximation as what is possible using an ideal non-realizable procedure which 'knows $\theta$'.

In short, orthonormal ridgelets, although not ridge functions, seem to have a role in linear approximation which is just as legitimate as the pure ridge functions $\psi^{+}_{j,k}(x_1 \cos \theta + x_2 \sin \theta)$.

1.8 Nonlinear Approximation by Ridgelets

We now consider nonlinear approximation to ridge functions by $N$-term ridgelet approximations. For this result, let $\eta_\delta(y, s) = y^1_{\{y > \delta\}}$ be a thresholding function with a "scaling" argument $s$, allowing for adjustment of the threshold. For a bounded function $f$, with $\tilde{N}(\delta) = \sum_{\lambda} 1\{\|f, \rho_\lambda\|_{L^\infty(D)} > \delta\}$ finite, set
\[
\tilde{f}_\delta = \sum_{\lambda} \eta_\delta(f, \rho_\lambda), \|\rho_\lambda\|_{L^\infty(D)} \rho_\lambda.
\]  
(1.10)

Under sufficient regularity of $f$, this is a finite sum of ridgelets, and so well-defined.

**Theorem 1.3** Let the ridge profile $r(t)$ belong to Besov space $\hat{B}^s_{p,p}(\mathbb{R})$, with $s = 1/p$ and $0 < p < 1$, and vanish at $\pm \infty$. Let $r_\theta(x)$ denote the corresponding ridge function of $x \in \mathbb{R}^2$. Define a sequence of approximants $(\tilde{r}_N)$ by letting $\tilde{f}_N(\delta)$ be the $\tilde{N}(\delta)$-term ridgelet approximation obtained from (1.10), with $f = r_\theta$. Then
\[
\|r_\theta - \tilde{r}_N\|_{L^\infty(D)} \leq C(r)N^{-(s-1)}, \quad N = 1, 2, \ldots.
\]  
(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 $\hat{B}^s_{p,p}$, 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 $L^\infty$ norm in $\mathbb{R}$ [11, 4]. 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 $r \in B_{p,p}^{s}[-d,d]$ alone would allow at best a conclusion of the form
\[
\|r - r_{N}\|_{L^\infty[-d,d]} \leq \text{Const} \cdot N^{-(s-1)}, \quad N = 1, 2, \ldots.
\] (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 $B_{p,p}^{s}[-d,d]$ for their approximation spaces [4].

In this light, Theorem 1.3 is rather interesting. It says that ridgelet thresholding does essentially as well, in a certain sense, as a kind of ideal pure ridge approximation which 'knows $\theta$', even though ridgelet thresholding does not 'know' the direction of the ridge. Indeed, an ideal ridge approximation could make use of $\theta$ as follows. With $r_{N}$ the $N$-term approximation to the ridge profile, just approximate $r_{\theta_{0}}$ using the ridge function $r_{N,\theta}$ in exactly the right direction, generated using the above approximate profile $r_{N}$ according to $r_{N}(x_{1}\cos\theta + x_{2}\sin\theta)$. Combining (1.12) with the analog for $r$ and $r_{N}$ of (1.9), this ideal scheme gives
\[
\|r_{\theta} - r_{N,\theta}\|_{L^\infty(D)} \leq \text{Const} \cdot N^{-(s-1)}, \quad N = 1, 2, \ldots,
\] (1.13)

and, from known results in best-$N$-term approximation, no other 'ideal ridge approximation scheme' is going to admit of substantially better estimates. By this we mean that 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.

Comparing (1.13) with (1.11) shows that, in a certain sense, an ideal approximation to a ridge function $r_{\theta}$ using $N$ pure ridge functions in the exact ridge direction does not do better than the $N$-term ridgelet approximation, which has no 'knowledge' of $\theta$ and does not depend in any way on an assumption that it be used on pure ridge functions.

In short, orthonormal ridgelets, although not ridge functions, seem to have a role in nonlinear approximation which is just as legitimate as approximation by pure ridge functions such as $\psi_{j,k}^{+}(x_{1}\cos\theta + x_{2}\sin\theta)$; and because the approximation is constructed using simple thresholding, the orthonormal ridgelet approach seems far better suited for possible algorithmic applications.
2 Preliminaries

2.1 Orthonormal Ridgelets and Radon Transform

We begin by briefly quoting some material from [5]. 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 \( L_{(\theta, t)} \), expressed using the Dirac mass \( \delta \) as

\[
(Rf)(t, \theta) = \int f(x) \delta(x_1 \cos \theta + x_2 \sin \theta - t) \, dx,
\]

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

\[
(Rf)(-t, \theta + \pi) = (Rf)(t, \theta).
\]

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

\[
F(-t, \theta + \pi) = F(t, \theta).
\]

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

\[
[F, G] = \frac{1}{4\pi} \int_{0}^{2\pi} \int_{-\infty}^{\infty} F(t, \theta) G(t, \theta) dt \, d\theta,
\]

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

\[
(P_{\mathcal{R}}F)(t, \theta) = (F(t, \theta) + F(-t, \theta + \pi))/2.
\]

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 \( \mathcal{R} \) consists of objects invariant under \( T \otimes S \); (2.3) can be rewritten \((T \otimes S)F = F\). In fact, \( P_{\mathcal{R}} = (I + T \otimes S)/2 \). Set now, for \( j, k \in \mathbb{Z} \), and \( i \geq i_0, \ell = 0, \ldots, 2^{i-1} - 1, \varepsilon \in \{0, 1\} \)

\[
W_{\lambda}(t, \theta) = P_{\mathcal{R}}(\psi_{j,k} \otimes w_{i,\ell}^{\varepsilon} ),
\]

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

\[
W_{\lambda}(t, \theta) = (\psi_{j,k}(t)w_{i,\ell}^{\varepsilon}(\theta) + \psi_{j,k}(-t)w_{i,\ell}^{\varepsilon}(\theta + \pi))/2.
\]

It was shown in [5] that the \( W_{\lambda} \) provide an orthobasis for \( \mathcal{R} \). In order to obtain orthonormality with respect to the scalar product \([ , ]\), a particular normalization was imposed. In that normalization, \( \|\psi_{j,k}\|_{L^2(\mathbb{R})} = \sqrt{2}, \)
and \( \|w_{\ell}^{\alpha}\|_{L^2[0,2\pi]} = 2\sqrt{\pi} \). 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 \(G \in \mathcal{R}\) and all sufficiently nice \(f \in L^2(dx)\),

\[
[Rf, G] = (f, R^+ G),
\]

which leads to

\[
(R^+ G)(x) = \frac{1}{4\pi} \int_{0}^{2\pi} G(x_1 \cos \theta + x_2 \sin \theta, \theta) \, d\theta.
\]

Define the Riesz order-1/2 fractional differentiation operator \(\Delta^+\) and also the order-1/2 fractional integration operator \(\Delta^-\) by the unified formula

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

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

Set now, for \(\lambda \in \Lambda\),

\[
\tau_{\lambda} = (\Delta^+ \otimes I)W_{\lambda}, \quad \lambda \in \Lambda.
\]

For example,

\[
\tau_{(j,k,i,\ell,1)} = (\Delta^+ \psi_{j,k} \otimes w_{1,\ell} + \Delta^+ T \psi_{j,k} \otimes S \omega_{1,\ell})/2.
\]

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

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

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 antipodally-symmetrized.
2.2 Special ridge functions

We now consider the relationship (1.3)-(1.4) between special ridge functions \( r^\gamma \) and ridgelets \( \rho_\lambda \). Formally, this is a simple matter, based immediately on (2.12) and the definition of the \( W_\lambda \). With \( \delta_\theta_0 \) denoting the Dirac mass at \( \theta_0 \in [0, 2\pi) \),

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

In a moment we will argue that

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

Combining (2.13) and (2.14), we have

\[
r^\gamma = R^+[(\Delta^+ \otimes I) \sum a_\lambda^j W_\lambda]
= R^+[(\sum a_\lambda^j (\Delta^+ \otimes I) W_\lambda]
= \sum a_\lambda^j R^+[(\Delta^+ \otimes I) W_\lambda]
= \sum a_\lambda^j \rho_\lambda,
\]

completing the formal argument for (1.4).

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

\[
\delta_{\theta_0} = \sum_{\ell=0}^{2^i-1} (\delta_{\theta_0}, \tilde{w}_{i,0,\ell} \tilde{w}_{i,0,\ell}^0 + \sum_{t=i_0}^{\infty} \sum_{\ell=0}^{2^i-1} (\delta_{\theta_0}, \tilde{w}_{i,\ell}^1 \tilde{w}_{i,\ell}^1);
\]

note here the expanded range of summation in \( \ell \), the range being \( 0 \leq \ell < 2^i \) rather than \( 0 \leq \ell < 2^{i-1} \) as in the definition of the set \( \Lambda \) from Section 1. In terms of the basis \( w_{i,\ell} \) without rescaling,

\[
4\pi \cdot \delta_{\theta_0} = \sum_{\ell=0}^{2^i-1} (\delta_{\theta_0}, w_{i,0,\ell} \tilde{w}_{i,0,\ell}^0 + \sum_{t=i_0}^{\infty} \sum_{\ell=0}^{2^i-1} (\delta_{\theta_0}, w_{i,\ell}^1 \tilde{w}_{i,\ell}^1). \tag{2.15}
\]
Hence, again in the sense of distributions,

\[
4\pi \cdot (\psi_{j,k} \otimes \delta_{\theta_0} + \psi_{j,1-k} \otimes \delta_{\theta_0 + \pi})/2
\]

\[
= \frac{1}{2} \sum_{\ell=0}^{2^{i_0-1}-1} \langle \delta_{\theta_0}, w_{i_0,\ell}^0 \rangle \psi_{j,k} \otimes w_{i_0,\ell}^0 + \frac{1}{2} \sum_{i=i_0}^{\infty} \sum_{\ell=0}^{2^{i-1}-1} \langle \delta_{\theta_0}, w_{i,\ell}^1 \rangle \psi_{j,k} \otimes w_{i,\ell}^1
\]

\[
+ \frac{1}{2} \sum_{\ell=0}^{2^{i_0-1}-1} \langle \delta_{\theta_0+\pi}, w_{i_0,\ell}^0 \rangle \psi_{j,1-k} \otimes w_{i_0,\ell}^0 + \frac{1}{2} \sum_{i=i_0}^{\infty} \sum_{\ell=0}^{2^{i-1}-1} \langle \delta_{\theta_0+\pi}, w_{i,\ell}^1 \rangle \psi_{j,1-k} \otimes w_{i,\ell}^1.
\]

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

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

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

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

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

\[
\sum_{\ell=0}^{2^{i_0-1}-1} \langle \delta_{\theta_0}, w_{i_0,\ell}^0 \rangle (\psi_{j,k} \otimes w_{i_0,\ell}^0 + \psi_{j,1-k} \otimes w_{i_0,\ell+2^{i_0-1}}^0)/2
\]

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

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

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

Obviously, similar relationships hold for sums involving terms \(w_{i,\ell}^1\) at indices \(i > i_0\), and so

\[
4\pi \cdot (\psi_{j,k} \otimes \delta_{\theta_0} + \psi_{j,1-k} \otimes \delta_{\theta_0 + \pi})/2
\]

\[
= \sum_{\ell=0}^{2^{i_0-1}-1} w_{i_0,\ell}(\theta_0) W_{(j,k;i_0,\ell,0)} + w_{i_0,\ell}(\theta_0 + \pi) W_{(j,1-k;i_0,\ell,0)}
\]

\[
+ \sum_{i=i_0}^{\infty} \sum_{\ell=0}^{2^{i-1}-1} w_{i,\ell}^1(\theta_0) W_{(j,k;i,\ell,1)} + w_{i,\ell}^1(\theta_0 + \pi) W_{(j,1-k;i,\ell,1)}.
\]

Comparing with the definition of \(W_\lambda\), 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 \( L^\infty \)-norm \( \|f\|_{\infty,-2\sigma} = \sup_{x \in \mathbb{R}^2} |f(x)|(1 + |x|^2)^{-\sigma}, \sigma \in \{3, 4, 5, \ldots\} \); this downweights \( x \)'s far from 0. We write \( L^\infty_{2\sigma} \) for the vector space of \( f \) with finite \( \|f\|_{\infty,-2\sigma} \). Let \( \Lambda^i \) denote the collection \( \{\lambda : j, i < i_1\} \) of indices \( \lambda \) that are index-limited to \( i, j \) coarser than \( i_1 \). Consider the indexlimited partial sum approximation \( r^{(i_1)} = \sum_{\Lambda^\cdot_i} a^\cdot_i \rho^\cdot_i \).

**Lemma 2.1** Let \( \sigma \in \{3, 4, 5, \ldots\} \). The identity

\[
r^\gamma = \sum_{\lambda} a^\cdot_i \rho^\cdot_i
\]

makes sense in \( L^\infty_{2\sigma} \); the approximation \( r^{(i_1)} \) 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 \( \partial \Lambda_{i_1, i_2} = \Lambda_{i_1}^\cdot \setminus \Lambda_{i_2}^\cdot, \ (i_2 > i_1) \),

\[
E_{i_1, i_2} = \left\| \sum_{\lambda \in \partial \Lambda_{i_1, i_2}} a^\cdot_i \rho^\cdot_i (1 + |\cdot|^2)^{-\sigma} \right\|_{L^\infty} 
\leq \sum_{\lambda \in \partial \Lambda_{i_1, i_2}} |a^\cdot_i| \|\rho^\cdot_i (1 + |\cdot|^2)^{-\sigma}\|_{L^\infty} 
\leq \sum_{\lambda \in \partial \Lambda_{i_1}^\cdot} |a^\cdot_i| \|\rho^\cdot_i (1 + |\cdot|^2)^{-\sigma}\|_{L^\infty} .
\]

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

\[
|a^\cdot_i| \leq C_m 2^{i/2}((1 + 2^i|\theta_0 - \theta_{i, \ell}))^{-m} + (1 + 2^i|\theta_0 + \pi - \theta_{i, \ell})^{-m}).
\]

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

\[
E_{i_1, i_2} \leq C(j_0) \sum_{i \geq i_1} 2^{-i(d + \frac{3}{2})} 2^{i/2} [((1 + 2^i|\theta_0 - \theta_{i, \ell}))^{-m} + (1 + 2^i|\theta_0 + \pi - \theta_{i, \ell})^{-m})]
= C'(j_0) \sum_{i \geq i_1} 2^{-id} \leq C''(j_0) 2^{-i_1d}.
\]

As \( E_{i_1, i_2} \leq C''(j_0) 2^{-i_1d} \) 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 \( r^\gamma \). The key observation is the following:

\[
\sum_{\ell=0}^{2^i - 1} (\tilde{w}_{i_0,1}^{0,\ell}, \delta_{\theta_0}) \tilde{w}_{i_0,1}^{0,\ell} + \sum_{i=i_0}^{i_1 - 1} \sum_{\ell=0}^{2^i - 1} (\tilde{w}_{i,1}^{1,\ell}, \delta_{\theta_0}) \tilde{w}_{i,1}^{1,\ell} = \sum_{\ell=0}^{2^{i_1} - 1} (\tilde{w}_{i_1,1}^{0,\ell}, \delta_{\theta_0}) \tilde{w}_{i_1,1}^{0,\ell},
\]

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 \( i_0 \leq i < i_1 \). Note that this relationship is habitually stated in terms of the orthonormal functions \( \tilde{w}_{i,1}^{\ell} \), but would remain equally valid in terms of \( w_{i,1}^{\ell} \). Let \( \tilde{P}_{i,1} \delta_{\theta_0} = \sum_{\ell=0}^{2^i - 1} (\tilde{w}_{i_0,1}^{0,\ell}, \delta_{\theta_0}) \tilde{w}_{i_0,1}^{0,\ell} \), and \( P_{i,1} \delta_{\theta_0} = \sum_{\ell=0}^{2^{i_1} - 1} (\psi_{i_1,1}^{0,\ell}, \delta_{\theta_0}) \psi_{i_1,1}^{0,\ell} \). Then \( P = 4\pi \cdot \tilde{P} \).

It follows immediately from the rewriting rule for \( P \) and the definition of \( a^\lambda_\gamma \) that

\[
\sum_{\lambda \in \Lambda^{i_1}} a^\lambda_\gamma \gamma = (\psi_{j_0,k_0}^+ \otimes P_{i_1} \delta_{\theta_0} + \psi_{j_0,1-k_0}^+ \otimes P_{i_1} \delta_{\theta_0 + \pi})/2
\]

\[
= 4\pi \cdot (\psi_{j_0,k_0}^+ \otimes \tilde{P}_{i_1} \delta_{\theta_0} + \psi_{j_0,1-k_0}^+ \otimes \tilde{P}_{i_1} \delta_{\theta_0 + \pi})/2.
\]

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

\[
r^{(i_1)} = \sum_{\lambda \in \Lambda^{i_1}} a^\lambda_\gamma \rho_\lambda = R^+ [\sum_{\lambda \in \Lambda^{i_1}} a^\lambda_\gamma \gamma]
\]

\[
= R^+ [ (\psi_{j_0,k_0}^+ \otimes P_{i_1} \delta_{\theta_0} + \psi_{j_0,1-k_0}^+ \otimes P_{i_1} \delta_{\theta_0 + \pi})/2] = 4\pi \cdot R^+ [ (\psi_{j_0,k_0}^+ \otimes \tilde{P}_{i_1} \delta_{\theta_0} + \psi_{j_0,1-k_0}^+ \otimes \tilde{P}_{i_1} \delta_{\theta_0 + \pi})/2].
\]

Put \( r_{j,k,\theta}(x) = \psi_{j,k}^+(x_1 \cos(\theta) + x_2 \sin(\theta)) \); fix \( x \) and view \( \theta \) 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 \([0,2\pi)\),

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

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

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

This pointwise convergence proves that \( r^{(i_1)} \rightarrow r^\gamma \) in \( L^\infty_{2\sigma} \) norm.

\[\Box\]

**Lemma 2.2** For \( \sigma \geq 3 \) and \( d = 1, \ldots, 2(\sigma - 2) \), for \( \epsilon = 1 \),

\[
\| \rho_\lambda(\cdot)(1 + |\cdot|^2)^{-\sigma} \|_{L^\infty} \leq C(j) 2^{-i(d+\frac{1}{2})}, \quad i \geq \max(i_0, j(\lambda)).
\]
PROOF. Let \( K(x) = (1 + |x|^2)^{-\sigma} \) then \( K \in L^1 \) provided \( \sigma > 1 \), and \( x_i^i K \in L^1 \) provided \( \sigma > 1 + \ell/2 \). It follows that \( \hat{K}(\xi) \in C^{2(\sigma-2)} \) and that \( \hat{K} \in C^\infty(\mathbb{R}^2 \setminus \{0\}) \). 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

\[
\|\rho_\lambda \cdot K\|_{L^\infty} \leq \|\hat{\rho}_\lambda \ast \hat{K}\|_{L^1}. \quad (2.17)
\]

Now as \( \hat{\rho}_\lambda \) is supported in an annulus \( \Gamma_j \), we have

\[
(\hat{\rho}_\lambda \ast \hat{K})(\xi) = \int_{\Gamma_j} \hat{\rho}_\lambda(\xi') \hat{K}(\xi - \xi') d\xi',
\]

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

\[
\int_{R_j}^{R_{j+1}} \int_0^{2\pi} (\hat{\psi}_{j,k}(r) \hat{w}_{i,k}(\theta) + \hat{\psi}_{j,1-k}(r) \hat{w}_{i,1-k}(\theta)) / 2 \cdot \hat{K}(\xi - (r \cos \theta, r \sin \theta)) d\theta r^{\frac{1}{2}} dr.
\]

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

\[
\hat{K}^{(d)}(r, \theta; \xi) = (\frac{\partial}{\partial \theta})^d \hat{K}(\xi - (r \cos \theta, r \sin \theta)),
\]

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

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

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

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

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

\[
\left\| \hat{K}^{(d)}(r, \theta; \xi) \right\|_{L^1(\theta \ d\theta)} \leq C_d \cdot C(j).
\]

Now for each fixed \( r > 0 \), if \( \varepsilon = 1 \), \( w_{i,\xi}^{\varepsilon} \) has \( d \)-fold primitive, and

\[
\int w_{i,\xi}^{\varepsilon}(\theta) \hat{K}(\xi - (r \cos \theta, r \sin \theta)) d\theta = (-1)^d \int (w_{i,\xi}^{\varepsilon})^{(d)}(\theta) \hat{K}^{(d)}(r, \theta; \xi) d\theta.
\]

Hence

\[
(\hat{\rho}_\lambda \ast \hat{K})(\xi) = \int_{R_j}^{R_{j+1}} \hat{\psi}_{j,k}(r) \left( \int_0^{2\pi} (w_{i,\xi}^{\varepsilon})^{(d)}(\theta) \hat{K}^{(d)}(r, \theta; \xi) d\theta \right) r^{\frac{1}{2}} dr + \int_{R_j}^{R_{j+1}} \hat{\psi}_{j,1-k}(r) \left( \int_0^{2\pi} (w_{i,\xi}^{\varepsilon})^{(d)}(\theta + \pi) \hat{K}^{(d)}(r, \theta; \xi) d\theta \right) r^{\frac{1}{2}} dr,
\]

16
and, by Minkowski,

\[
\| \hat{\rho} \ast \hat{K} \|_{L^1} \leq \sup_{\xi \in \Gamma_j} \| \hat{K}^{(d)}(\cdot, \cdot; \xi) \|_{L^1(dr \, d\theta)} \\
\times \left( \int_{R_j} \int_0^{2\pi} |\psi_{j,k}(r)| r^{-\frac{1}{2}} \left| (w_{\ell,\ell}^{(d+\frac{1}{2})})^{-d}(\theta) \right| \, d\theta dr \\
+ \int_{R_j} \int_0^{2\pi} |\hat{\psi}_{j,1-k}(r)| r^{-\frac{1}{2}} \left| (w_{\ell,\ell}^{(d+\frac{1}{2})})^{-d}(\theta + \pi) \right| \, d\theta dr \right) / 2 \\
= C_d \cdot \int_0^{2\pi} \left| (w_{\ell,\ell}^{(d+\frac{1}{2})})^{-d}(\theta) \right| d\theta \\
= C_d \cdot 2^{-i(d+\frac{1}{2})}. \quad \diamond
\]

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 \( r_{\theta} \) with profile \( r \in B_{1,1}^1(\mathbb{R}) \).

**Lemma 3.1** Let \( r(t) \) belong to the Bump algebra — Besov space \( B_{1,1}^1(\mathbb{R}) \) — 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 \( L^\infty(\mathbb{R}) \) 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 \( r(t) \in B_{1,1}^s(\mathbb{R}), \ s \geq 1, \) and vanish at \( \pm \infty \). There exist coefficients \((c_{j,k} : j, k \in \mathbb{Z})\) so that

\[
r = \sum c_{j,k} \psi_{j,k}^{(+)}
\]

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

\[
\sum |c_{j,k}| 2^{js} < C \cdot \| r \|_{B_{1,1}^s(\mathbb{R})}. \quad (3.1)
\]

**Proof of Theorem 1.1.** Let \( \gamma = (j_0, k_0, \theta_0) \) and \( a_{\lambda}^{(\gamma)} \) as in (1.3). Let \( r^{(i_1)} = \sum a_{\lambda}^{(\gamma)} \rho_{\lambda} \). 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 $\mathbb{R}^2$ 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 $(\rho_\lambda, \cdot)$ defines a bounded linear functional on $L^\infty_{2\sigma}$. Hence

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

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

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

It follows that

$$\langle r^\gamma, \rho_\lambda \rangle = a_\lambda^{(i)} \quad \forall \ \lambda \in \Lambda.$$  \hspace{1cm} (3.2)

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

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

which proves (1.7) in the special case where $r_\lambda$ 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 $r_N, \theta_0 = \sum_{n=1}^N r_N$ 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):

$$\langle r_N, \theta_0, \rho_\lambda \rangle = \left( \sum_{n=1}^N c_n r_N, \rho_\lambda \right) = \sum_{n=1}^N c_n \cdot \langle r_N, \rho_\lambda \rangle = \sum_{n=1}^N c_n \cdot \left( \langle \psi^+_{j_n,k_n}, \psi^-_{j,k} \rangle \cdot w^\xi_{\ell,\ell}(\theta_0) + \langle \psi^+_{j_n,k_n}, \psi^-_{j,1-k} \rangle \cdot w^\xi_{\ell,\ell}(\theta_0 + \pi) \right) / 2$$

$$= \left( \langle \sum_{n=1}^N c_n \psi^+_{j_n,k_n}, \psi^-_{j,k} \rangle \cdot w^\xi_{\ell,\ell}(\theta_0) + \langle \sum_{n=1}^N c_n \psi^+_{j_n,k_n}, \psi^-_{j,1-k} \rangle \cdot w^\xi_{\ell,\ell}(\theta_0 + \pi) \right) / 2$$

$$= \langle r_N, \psi^-_{j,k} \rangle \cdot w^\xi_{\ell,\ell}(\theta_0) + \langle r_N, \psi^-_{j,1-k} \rangle \cdot w^\xi_{\ell,\ell}(\theta_0 + \pi) / 2.$$

But

$$r_N, \theta_0 \to r_{\theta_0} \text{ in } L^\infty(\mathbb{R}^2) \text{ and } r_N \to r \text{ in } L^\infty(\mathbb{R}) \text{ as } N \to \infty.$$  \hspace{1cm}

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

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

So (1.7) follows as claimed.
4 Synthesis of Ridge Functions

We now consider the formal association

\[ r_{\theta_0} \sim \sum_{\Lambda} \langle r_{\theta_0}, \rho_{\lambda} \rangle \rho_{\lambda} \]

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 \( D = \{ x \in \mathbb{R}^2 : |x| \leq 1 \} \), and consider the approximation operator

\[ A^i f = \sum_{\Lambda \in \Lambda^i} \langle f, \rho_{\lambda} \rangle \rho_{\lambda} \]

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

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

**Proof.** Decompose \( \Lambda^i \) into layers \( L^i = \Lambda^i \setminus \Lambda^{i-1} \); i.e. for \( i_1 > i_0 \) put

\[ \Lambda^i = \Lambda^{i_0} \cup L^{i_0+1} \cup \ldots \cup L^{i_1} \]

Let \( B^i \) be defined formally by

\[ B^i f = \sum_{L^i} \langle f, \rho_{\lambda} \rangle \rho_{\lambda} \]

Then formally

\[ A^i = A^{i_0} + B^{i_0+1} + \ldots + B^{i_1} \]

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 \( m = 1, 2, \ldots \)

\[ |\langle r^\gamma, \rho_{\lambda} \rangle| \leq C_m \cdot \delta_{j_0 \epsilon} \cdot (\delta_{k_0,1-k}) \cdot 2^{i/2} \cdot [1 + 2^i |\theta_0 - \theta_i, \epsilon|]^{-m} + (1 + 2^i |\theta_0 + \pi - \theta_i, \epsilon|)^{-m} \]

and as Lemma 4.2 below gives \( \| \rho_{\lambda} \|_{L^\infty(D)} \leq C_d \cdot 2^{i/2} \cdot 2^{-i(1-f)} \cdot d \)

\[ \sum_{L^i} |\langle r^\gamma, \rho_{\lambda} \rangle| \| \rho_{\lambda} \|_{L^\infty(D)} \leq 2^{i_0} \cdot 2^{-i_0(d-1/2)} \cdot C \]

(4.1)

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

19
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 $\sum |c_n| 2^{j_n} \leq C$, 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

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

Clearly each sum corresponding to $B^1(t^{(N)}_\theta)$ is norm-convergent. Moreover, by (4.1),

$$\|B^1(r^{(N_2)}_\theta - r^{(N_1)}_\theta)\|_{L^\infty} \leq \sum_{n=N_1}^{N_2} \|B^1(c_n r^{\gamma_n})\|_{L^\infty} \leq C \cdot \sum_{n=N_1}^{N_2} 2^{j_n} |c_n|.$$

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

$$\|B^1(r^{(N_2)}_\theta - r^{(N_1)}_\theta)\|_{L^\infty} \leq \sum_{n=N_1}^{\infty} 2^{j_n} |c_n| \leq C \|r\|_{\dot{B}^1_{1,1}}. \quad (4.2)$$

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

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

We now turn to the 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 $r^{(i_1)} = \sum_{j \leq i_1} c_{j,k} \psi_{j,k}^+$ be an approximation to the ridge profile generated using only terms $\psi_{j,k}^+$ at indices $j \leq i_1$. Note that the index-limiting approximation operator $A^{i_1}$ behaves the same on the ridge function $r_\theta$ as on the ridge function with the approximate profile $r^{(i_1)}_\theta$:

$$A^{i_1} r^{(i_1)}_\theta = A^{i_1} r_\theta. \quad (4.4)$$

The triangle inequality gives

$$\|r_\theta - A^{i_1} r_\theta\|_{L^\infty(D)} \leq \|r_\theta - r^{(i_1)}_\theta\|_{L^\infty(D)} + \|r^{(i_1)}_\theta - A^{i_1} r^{(i_1)}_\theta\|_{L^\infty(D)} \quad (4.5)$$
Now as \( r \in \hat{B}_{t_1} \), (3.1) can be used, along with rapid decay of \( w^r_{i,t} \), (1.9) and (4.8), giving
\[
\| r_{\theta_0} - r^{(t_1)}_{\theta_0} \|_{L^\infty(D)} \leq \sum_{j > t_1} \sum_{k} |c_{j,k}| \| \psi_{j,k}^+ \|_{L^\infty[-d,d]} \\
\leq C \sum_{j > t_1} \sum_{k} |c_{j,k}| 2^j \\
\leq C \sum_{j > t_1} 2^{-(s-1)j} \left( \sum_{k} |c_{j,k}| 2^{js} \right) \\
\leq C \cdot 2^{-(s-1)j_1 \sum_{j > j_1} \sum_{k} |c_{j,k}| 2^{js}} \\
\leq C' \cdot 2^{-(s-1)j_1} \| r \|_{\hat{B}_{t_1}(R)} .
\]  
(4.6)

Meanwhile
\[
\| r^{(t_1)}_{\theta_0} - A^{t_1} r^{(t_1)}_{\theta_0} \|_{L^\infty(D)} \leq \sum_{i \geq t_1} |\langle r^{(t_1)}_{\theta_0}, \rho_\lambda \rangle| \| (I - A^{t_1}) \rho_\lambda \|_{L^\infty(D)} .
\]  
(4.7)

Now by Theorem 1.1,
\[
|\langle r^{(t_1)}_{\theta_0}, \rho_\lambda \rangle| \leq \sum_j |c_j| \cdot |\langle e^\gamma, \rho_\lambda \rangle| \cdot 1_{\{j \leq j_1\}} \\
\leq \sum_{j \geq j_0, \kappa_0} |c_{j_0, \kappa_0}| \left( \delta_{j_0, \kappa_0} \cdot (\delta_{k, (1-k_0)} \cdot (|w_{i,t}(\theta_0)| + |w_{i,t}(\theta_0 + \pi)|)) \right) \cdot 1_{\{j \leq j_1\}} \\
= (|c_{j,1-k_1}|) \cdot (|w_{i,t}(\theta_0)| + |w_{i,t}(\theta_0 + \pi)|) \cdot 1_{\{j \leq j_1\}} .
\]

Also
\[
(I - A^{t_1}) \rho_\lambda = \begin{cases} 0 & j(\lambda) \leq i(\lambda) \leq j_1 \\ -\rho_\lambda & j(\lambda) \leq j_1 < i(\lambda) \end{cases} .
\]

Lemma 4.2 below gives for \( d > 0 \)
\[
\| (I - A^{t_1}) \rho_\lambda \|_{L^\infty(D)} \leq \begin{cases} 2^{j(\lambda)/2} 2^{-(s)j(\lambda)} \delta_{d} & j(\lambda) \leq i_1 < i(\lambda) \\ 0 & j(\lambda) \leq i(\lambda) \leq i_1 . \end{cases}
\]

21
Hence the right-hand side of (4.7) is upper-bounded by, for \(d > s\),

\[
C \cdot \sum_{j \leq i_1} \sum_{k} |c_{j,k}| 2^{j/2} \sum_{i > i_1} 2^{i/2} \ 2^{-(i-j)d} \\
\times \left( \sum_{\ell} (1 + 2^{j} |\theta_0 - \theta_{i,\ell}|)^{-m} + \sum_{\ell} (1 + 2^{j} |\theta_0 + \pi - \theta_{i,\ell}|)^{-m} \right) \\
\leq C \sum_{j \leq i_1} \sum_{k} |c_{j,k}| 2^{j/2} \ 2^{-(i_1-j)d} 2^{i_1/2} \\
\leq C \sum_{j \leq i_1} 2^{-(i_1-j)d} 2^{i_1/2} 2^{j/2} \ 2^{-(s-j)\sum_{k} |c_{j,k}| 2^{js}} \\
= C 2^{-i_1(d-1/2)} \sum_{j \leq i_1} 2^{j(d+1/2-s)} \sum_{k} |c_{j,k}| 2^{js} \\
\leq C 2^{-i_1(d-1/2)} \sum_{j \leq i_1} 2^{i_1(d+1/2-s)} \frac{\|r\|_{\mathcal{B}_{1,1}(R)}}{2^{i_1(s-1)}} \\
\leq C 2^{-i_1(s-1)} \frac{\|r\|_{\mathcal{B}_{1,1}(R)}}{2^{i_1(s-1)}}.
\]

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

\[\diamondsuit \]

**Lemma 4.2** For \(d > 0\),

\[
\|\rho_{\lambda}\|_{L^\infty(D)} \leq C_d \cdot 2^{j/2} \ 2^{-(t-j)d}, \quad \lambda \in \Lambda, \ \epsilon(\lambda) = 1. \tag{4.8}
\]

**Proof.** From (1.2),

\[
\rho_{\lambda}(x) = \frac{1}{4\pi} \int \psi_{j,k}^+(x_1 \cos \theta + x_2 \sin \theta)w_{i,\ell}^+(\theta)d\theta.
\]

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

\[
\left(\frac{\partial}{\partial \theta}\right)^d f_x(\theta) = \sum_{\ell=0}^{d} \left(\frac{\partial}{\partial \ell}\right)^\ell \psi_{j,k}^+(t) \bigg|_{t=t(\theta)} P_{d,\ell}(t(\theta), \ldots, t^{(m)}(\theta))
\]

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

\[
\left|\left(\frac{\partial}{\partial \ell}\right)^\ell \psi_{j,k}^+(t) \right|_{t=t(\theta)} \leq 2^{j\ell} 2^{j\ell} C_{m,\ell}(1 + 2^{j}\ell |t(\theta) - t_{j,k}|)^{-m}.
\]

Hence

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

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

\[
\left\| (w_{i,\ell}^1)^{(-d)} \right\|_{L^2(d\theta)} \leq C 2^{-id},
\]

22
\[ |\rho_\lambda(x)| = c \left| \int \left[ \frac{\partial}{\partial \theta} \right]^d f_\lambda \left[ \frac{\partial}{\partial \theta} \right]^{-d} w_{i,k} \right| d\theta \leq C_d 2^{j/2} 2^{-(i-j)d}, \quad i > j, \quad d = 1, 2, \ldots \]

5 Nonlinear Approximation of Ridge Functions

We now turn to the proof of Theorem 1.3. A simple adaptation of the earlier lemma concerning \( \tilde{B}_{i,1} \) will prove

**Lemma 5.1** Let \( r \in \tilde{B}_{s,p}^s(\mathbb{R}) \), 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 \( L^\infty \), and for \( C < \infty \),

\[
\sum_{j,k} |c_{j,k}|^{2ips} \leq C^p. \quad (5.1)
\]

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 interpolation theory – probably contained in Bergh-Lofstrom (1976); it is provable by elementary means. We omit the argument.

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

\[
S_N \leq C_p \cdot V \cdot N^{1-1/p}, \quad N = 1, 2, 3, \ldots \quad (5.2)
\]

So, 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,

\[
r_{\theta_0} = \sum_{j_0,k_0} c_{\gamma} r^\gamma, \quad (5.3)
\]

where, as \( r \in \tilde{B}_{s,p}^s(\mathbb{R}) \) for \( s = 1/p \),

\[
\sum_{j_0,k_0} |c_{\gamma}|^p 2^{|\gamma|} \leq C^p \quad (5.4)
\]
for some $C > 0$.

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

$$x^\gamma_\lambda = (r^\gamma, \rho\lambda) \cdot \|\rho\lambda\|_{L^\infty(D)}, \quad \lambda \in \Lambda;$$

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

$$\|x^\gamma\|_{\ell^p} \leq C_p \cdot 2^{j_0}, \quad (5.5)$$

where $C_p$ does not depend on $\gamma$; this justifies the appellation 'molecular'. Now the $\ell^p$-quasi norm obeys the $p$-triangle inequality

$$\|x + y\|_{\ell^p} \leq \|x\|_{\ell^p}^p + \|y\|_{\ell^p}^p \quad (5.6)$$

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 $r_{\theta_0}$. The array $x$ obeys, by (5.3)–(5.4), the molecular decomposition

$$x = \sum_{j_0, k_0} c_\gamma x^\gamma_\lambda. \quad (5.7)$$

Combining this with (5.5) and (5.6),

$$\|x\|_{\ell^p}^p = \|\sum_{j_0, k_0} c_\gamma x^\gamma_\lambda\|_{\ell^p}^p \leq \sum_{j_0, k_0} |c_\gamma|^p \|x^\gamma\|_{\ell^p}^p \leq C \sum_{j_0, k_0} |c_\gamma|^p 2^{j_0p} \leq C' \cdot \|r\|_{B^p_{p,p}(\mathbb{R})}^p. \quad (5.8)$$

Now consider

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

24
We have, with \( \Lambda^\delta = \{ \lambda : |(r_{\theta_0}, \rho_\lambda)| \cdot \|\rho_\lambda\|_{L^\infty(D)} < \delta \}, \)

\[
\|r_{\theta_0} - \tilde{r}^{(\delta)}\|_{L^\infty(D)} = \| \sum_{\Lambda^\delta} (r_{\theta_0}, \rho_\lambda) \rho_\lambda\|_{L^\infty(D)} \leq \sum_{\Lambda^\delta} |(r_{\theta_0}, \rho_\lambda)| \cdot \|\rho_\lambda\|_{L^\infty(D)} \tag{5.9}
\]

Let \( v_1, v_2, \ldots \) be an enumeration of the entries \( |(r_{\theta_0}, \rho_\lambda)| \cdot \|\rho_\lambda\|_{L^\infty(D)} \) in nonincreasing order. There is an \( N = N(\delta) \) so that the sum on the right of (5.9) 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 \( C \cdot \|r\|_{L_{\ell,p}(R)} \) according to (5.8), we have by (5.2)

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

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

\[
\|x^\gamma\|_{\ell^p} \leq C_p \cdot 2^{j_0}, \tag{5.10}
\]

where \( C_p \) is independent of \( (j_0, k_0, \theta_0) \).

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

\[
|a^\gamma_\lambda| \leq C_m \cdot 2^{i/2} \cdot ((1 + 2^i|\theta_0 - \theta_i, \ell|)^{-m} + (1 + 2^i|\theta_0 + \pi - \theta_i, \ell|)^{-m}), \quad \forall i, \ell \tag{5.11}
\]

for each \( m > 1 \). Moreover, by Lemma 4.2,

\[
\|\rho_\lambda\|_{L^\infty(D)} \leq C_d \cdot 2^{j_0/2} \cdot 2^{-(i-j_0)d}, \tag{5.12}
\]

for each \( d > 0 \). Set now, for fixed \( m \),

\[
u^\gamma(t) = (1 + 2^i|t - \theta_i, \ell|)^{-m},
\]

and note that by (5.11), \( |a^\gamma_\lambda| \leq C_m \cdot 2^{i/2} \cdot (u^\gamma_0(\theta_0) + u^\gamma_0(\theta_0 + \pi)). \) Now for \( mp > 1 \), and \( t \in [0, 2\pi), \)

\[
\|(u^\gamma_\ell(t))_\ell\|_{\ell^p} \leq C_{mp} < \infty. \tag{5.13}
\]

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

\[
u^h = (\nu^h_\ell : \ell = 0, \ldots, 2^{j_0+h-1}-1),
\]

25
and note that the nonzero entries in $x^\gamma$ 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 $x^\lambda_i$ such that $i = j_0 + h, k = k_0, \ell = 0, \ldots, 2^{j_0 + h - 1} - 1$. Now combining (5.11) and (5.12), we have, with $A_h = C_d \cdot 2^{j_0/2} \cdot 2^{-h \delta}$,

$$\|v^h\|_{P^p} \leq A_h^p \cdot (\|u^{j_0+h}(\theta_0)\|_{P^p} + \|u^{j_0+h}(\theta_0 + \pi)\|_{P^p}) \cdot 2^{(j_0+h)p/2} \leq c_{m,p} A_h^p 2^{(j_0+h)p/2}.$$  

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

$$\|v\|_{P^p} = \|v^0\|_{P^p} + \|v^1\|_{P^p}.$$  

Using this and picking $d > 1/2$, so that $\sum_{h\geq 0} (2^{-h \delta} \pi^p)2^{hp/2} \leq C_{d,p}$, we have

$$\|(v^0, v^1, \ldots)\|_{P^p} = \sum_{h\geq 0} \|v^h\|_{P^p} \leq c_{m,p} \sum_{h\geq 0} A_h^p \cdot 2^{(j_0+h)p/2} \leq C' 2^{j_0p},$$

and so (5.10) follows.

6 Appendix

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

$$\|\sum c_{j,k} \psi^+_{j,k}\|_{L^\infty} \leq \sum |c_{j,k}| \|\psi^+_{j,k}\|_{L^\infty} \leq C \sum |c_{j,k}| 2^j < \infty.$$  

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

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

then we have

$$\|r - r_N\|_{L^\infty} \leq \sum_{n=1}^N a_n.$$  

By (6.1), $r \in \hat{B}_{1,1}^1(\mathbb{R})$ implies $\sum_{n=1}^\infty a_n < \infty$, so $\sum_{N=1}^\infty a_n \to 0$ as $N \to \infty$.  

26
Proof of Lemma 3.2. We have, from $\Delta^+\Delta^+ = \text{Id}$ on $\text{Dom}(\Delta^+) \cap \text{Dom}(\Delta^-)$, from self-adjointness of $\Delta^\pm$, and from $\psi_{j,k} \in \text{Dom}(\Delta^+) \cap \text{Dom}(\Delta^-)$,

\[
\langle \psi_{j',k'}, \psi_{j,k} \rangle = (\Delta^+ \psi_{j',k'}, \Delta^- \psi_{j,k}) = (\Delta^- \Delta^+ \psi_{j',k'}, \psi_{j,k}) = (\psi_{j',k'}, \psi_{j,k}) = \delta_{j,j'} \cdot \delta_{k',k} .
\]

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 $\nu(\omega)$ 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 $\Delta_j^+$ and $\Delta_j^-$ by

\[
(\Delta_j^+ f)(t) = \int \hat{f}(\omega) e^{i\omega t} |\omega|^{1/2} \nu(2^{-j} \omega) d\omega .
\]

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

\[
\Delta_j^\pm \psi_{j',k'} = \Delta_j^\pm \psi_{j',k'} , \quad j' \in \{j-1, j, j+1\} , \quad k \in \mathbb{Z} .
\]

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

\[
\Delta_j^- \psi_{j',k'} = \sum_{j'=j'-1}^{j'+1} \sum_k \langle \Delta_j^- \psi_{j',k'}, \psi_{j,k} \rangle \psi_{j,k} .
\]

It follows from self-adjointness of $\Delta_j^-$,

\[
\langle \psi_{j',k'}, \psi_{j,k} \rangle = \langle \Delta_j^- \psi_{j',k'}, \psi_{j,k} \rangle = \langle \psi_{j',k'}, \Delta_j^- \psi_{j,k} \rangle = \langle \psi_{j',k'}, \psi_{j,k}^- \rangle , \quad j \in \{j' - 1, j', j' + 1\} .
\]
So we may write
\[
\psi_{j',k'}^{+} = \sum_{j=j'-1}^{j'+1} \sum_{k} \langle \psi_{j',k'}, \psi_{j,k}^{-} \rangle \psi_{j,k}.
\]

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
\[
A(j', k'; j, k) = \int \psi_{j',k'}(t) \psi_{j,k}(t) \, dt \\
= 2^{-j'/2} \int \psi_{0,k'}(2^{j'} t) \psi_{j-j',k}(2^{j'} t) 2^{j'} \, dt \\
= 2^{-j'/2} \langle \psi_{0,k'}, \psi_{j-j',k} \rangle \\
= 2^{-j'/2} A(0, k'; j - j', k).
\]
(6.3)

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
\[
r(t) = \sum \alpha_{j',k'} \psi_{j',k'} \\
= \sum \alpha_{j',k'} \sum A(j'k'; j, k) \psi_{j,k}^{+} \\
= \sum \psi_{j,k}^{+} \sum A(j'k'; j, k) \alpha_{j',k'}.
\]

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

\[ \sum_{k'} |A(0,k',h,k)| \leq C \quad \forall k, \forall h = -1,0,1, \quad (6.4) \]

and we have

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

In short, for \( s \geq 1 \),

\[ \sum_{j,k} 2^{js} |a_{j,k}| \leq C \sum_{j',k'} |\alpha_{j'k'}| 2^{(s-1)/2} . \quad \diamond \]

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

\[ T\psi_{0,0} = \sum_{j=-\infty}^{\infty} \sum_{k} \langle T\psi_{0,0}, \psi_{j,k} \rangle \psi_{j,k} \ . \]

**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 write \( \hat{T}\psi(\omega) = \tau(\omega)\hat{\psi}(\omega) \) with \( \tau(\omega) \) an essentially bounded function, and so the support of \( \hat{T}\psi(\omega) \) is contained in the support of \( \hat{\psi}(\omega) \). Thus \( \hat{T}\psi_{0,0}(\omega) \) 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 \( j > 1 \).

\[ \diamond \]

**Lemma 6.2.** For \( 0 < p \leq 1 \),

\[ \sum_{k'} |A(0,k',h,k)|^p \leq C_p \quad \forall k, \forall h = -1,0,1, \quad (6.5) \]

**Proof.** Writing \( \psi \) for \( \psi_{0,0} \) and passing to the frequency domain,

\[ A(0,k',h,k) = \frac{1}{2\pi} \int \hat{\psi}(\omega) e^{-i\omega k'} \hat{\psi}(\omega/2^h)^* e^{i\omega k'/2^h} 2^{-h/2}d\omega \]

\[ = \frac{1}{2\pi} \int \hat{\psi}(\omega) \hat{\psi}(\omega/2^h)^* 2^{-h/2} e^{-i\omega (k'-k/2^h)}d\omega \]

\[ = \Psi_h(k/2^h - k'), \quad \text{say} , \]
where
\[ \hat{V}_h(\omega) \equiv \hat{\psi}(\omega)\hat{\psi}(\omega/2^h)^* 2^{-h/2}. \]

Now \( \hat{V}_h \) is \( C^\infty \) and of compact support. Hence \( \psi_h \) is of rapid decay, and so for each \( m > 0 \) we have \( C_m \) with
\[ |\psi_h(t)| \leq C_m \cdot (1 + |t|)^{-m} \quad \forall t \in \mathbb{R}. \]

Now picking \( mp > 1 \),
\[ \sum_{k'} |A(0, k', h, k)|^p = \sum_{k'} |\psi_h(k/2^h - k')|^p \leq C_m^p \sum_{k'} (1 + |k/2^h - k'|)^{-mp} \leq C_{m,p} < \infty. \quad \Diamond \]

References


